ESTATÍSTICA APLICADA
à Administração e Economia

2ª edição

David R. Anderson
Dennis J. Sweeney
Thomas A. Williams

Alexei Magalhães Veneziani

Pode Cortar

Estatística Aplicada à Administração e Economia

Dados Internacionais de Catalogação na Publicação (CIP)
(Câmara Brasileira do Livro, SP, Brasil)

Anderson, David R.
Estatística aplicada à administração e economia / David R. Anderson, Dennis J. Sweeney, Thomas A. Williams ; (José Carlos Barbosa dos Santos - ERJ Composição Editorial e Artes Gráficas Ltda. ; revisão técnica Petrônio Garcia Martins) -- 2. ed. -- São Paulo: Cengage Learning, 2011.

Título original: Essentials of statistics for business and economics
4. reimpr. da 2. ed. de 2007.
Bibliografia
ISBN 978-85-221-0521-2

1. Ciências sociais - Métodos estatísticos 2. Economia - Métodos estatísticos 3. Estatísticas - Problemas e exercícios 4. Estatística comercial I. Sweeney, Dennis J. II. Williams, Thomas A. III Título.

06-7974 CDD-519.5

Índice para catálogo sistemático:

1. Estatística aplicada 519.5

Estatística Aplicada à Administração e Economia

2ª Edição

David R. Anderson

University of Cincinnati

Dennis J. Sweeney

University of Cincinnati

Thomas A. Williams

Rochester Institute of Technology

Tradução

José Carlos Barbosa dos Santos – ERJ Composição Editorial e Artes Gráficas Ltda.

Revisão Técnica

Petrônio Garcia Martins

Mestre em Engenharia de Produção pela Escola Politécnica da USP

Professor da FEI e da FMU

Austrália • Brasil • Japão • Coréia • México • Cingapura • Espanha • Reino Unido • Estados Unidos

Estatística Aplicada à Administração e Economia
2ª edição

David R. Anderson / Dennis J. Sweeney / Thomas A. Williams

Gerente Editorial: Patricia La Rosa

Editora de Desenvolvimento: Danielle Mendes Sales

Supervisor de Produção Editorial: Fábio Gonçalves

Produtora Editorial: Gabriela Trevisan

Supervisora de Produção Gráfica: Fabiana Alencar Albuquerque

Título do Original em Inglês: Essentials of Statistics for Bussiness and Economics, 4th edition

ISBN Original: 0-324-22322-6

Tradução: José Carlos Barbosa dos Santos – ERJ Composição Editorial e Artes Gráficas Ltda.

Revisão Técnica: Petrônio Garcia Martins

Copidesque: Maria Alice da Costa

Revisão: Mônica Cavalcante Di Giacomo e Sandra Garcia Cortés

Composição: Cia. Editorial

Capa: Fz.Dáblio Design Studio

ISBN-10: 85-221-0521-9
ISBN-13: 978-85-221-0521-2

Cengage Learning
Condomínio E-Business Park
Rua Werner Siemens, 111 – Prédio 20 – Espaço 04
Lapa de Baixo – CEP 05069-900 – São Paulo – SP
Tel.: (11) 3665-9900 – Fax: (11) 3665-9901
SAC: 0800 11 19 39

Para suas soluções de curso e aprendizado, visite
www.cengage.com.br

Impresso no Brasil.
Printed in Brazil.
2 3 4 5 6   11 10 09 08 07

Dedicado a
Marcia, Cherri e Robbie

Apresentação

O objetivo do livro *Estatística Aplicada à Administração e Economia*, 2ª edição, é oferecer aos alunos, principalmente aqueles das áreas de Administração e Economia, a introdução conceitual do campo da Estatística e suas muitas aplicações. O texto é orientado e foi escrito tendo em mente as necessidades do aluno não-matemático; o único pré-requisito matemático exigido é o conhecimento de álgebra.

As aplicações de análise de dados e metodologia estatística são parte integrante da organização e apresentação do conteúdo. A discussão e o desenvolvimento de cada técnica são apresentados num conjunto aplicativo, com os resultados estatísticos fornecendo critérios para decisões e soluções de problemas.

Apesar de o livro ser orientado para aplicações, tivemos o cuidado de proporcionar um desenvolvimento metodológico correto e usar a notação geralmente aceita para o tópico em discussão. Assim, os alunos descobrirão que o texto oferece boa preparação para o estudo de material estatístico mais avançado. Uma bibliografia revisada e atualizada para orientar estudos adicionais foi incluída como apêndice.

Mudanças Nesta Edição

Agradecemos a aceitação e resposta positiva às edições anteriores de *Estatística Aplicada à Administração e Economia*. Conseqüentemente, ao introduzir modificações nesta nova edição, mantivemos o estilo da apresentação e legibilidade daquelas edições. As mudanças significativas nesta edição estão resumidas a seguir.

Revisões do Conteúdo

A seguinte lista resume as revisões de conteúdo selecionadas para esta edição.

- **Estimação por intervalo**: Nas edições anteriores, seguimos a abordagem de amostra grande/amostra pequena para estimação por intervalo da média da população no Capítulo 8. Na nova edição, apresentamos a estimação por intervalo usando os paradigmas σ conhecido e σ desconhecido. A distribuição normal padrão é empregada em todos os casos em que o desvio padrão da população possa ser conhecido. A distribuição t é usada em todos os casos em que o desvio padrão da população é estimado pelo desvio padrão da amostra. Essa abordagem simplifica a metodologia para o aluno e é consistente com os procedimentos baseados em computador oferecidos pelo Minitab e pelo Excel. No caso de σ desconhecido, a nova abordagem fornece resultados relativamente melhores que a aproximação anterior da amostra grande. Uma tabela da distribuição t com até 100 graus de liberdade foi incluída sob a

designação de Tabela 2 no Apêndice B. Essa mudança leva aos testes de hipóteses sobre a média da população no Capítulo 9 e a inferências estatísticas sobre duas médias da população no Capítulo 10.

- **Testes de hipóteses usando valores *p***: Outra mudança na edição nova é a ênfase no uso de valores *p* para teste de hipóteses. Com a utilização de pacotes de programas estatísticos para análise de dados cada vez mais difundido, os valores *p* são claramente preferidos à abordagem tradicional de teste estatístico e região de rejeição. Como conseqüência, atualmente se usam valores *p* como o método principal para aplicações de teste de hipóteses nos Capítulos 9 a 13.
- **Procedimento novo para inferências de duas amostras**: Oferecemos nova metodologia para inferências sobre médias de duas populações quando os desvios padrão da população forem desconhecidos. A metodologia é baseada na distribuição *t* e é bem mais genérica porque pode ser aplicada, sendo ou não iguais às variâncias da população. O aluno não precisa mais considerar a hipótese de igualdade da variância da população e efetuar o cálculo da variância agrupada.
- **Estatística descritiva**: Foram adicionadas seções novas nos Capítulos 2 e 3 sobre o formato das distribuições. A assimetria foi introduzida como medida importante do formato da distribuição. Nos capítulos finais mencionamos agora a necessidade de tamanhos de amostras maiores para estimação por intervalo e teste de hipóteses nas aplicações que envolvem população assimétrica. O material sobre tabulação cruzada foi ampliado para incluir mais discussão acerca de distribuições de porcentagens. O paradoxo de Simpson é usado para indicar uma fonte de conclusões potencialmente errôneas ao trabalhar com tabulações cruzadas.
- **Distribuições de probabilidade**: Foi acrescentada uma discussão sobre média, variância e desvio padrão para as distribuições de Poisson e hipergeométrica no Capítulo 5 e para a distribuição exponencial no Capítulo 6. Esse capítulo também tem uma nova seção sobre aproximação normal da probabilidade binomial.

Exemplos e Exercícios Novos Baseados em Dados Reais

Acrescentamos aproximadamente 200 exemplos e exercícios novos baseados em dados reais e fontes de referências recentes de informações estatísticas. Usamos o *Wall Street Journal*, o *USA Today*, a *Fortune*, a *Barron's* e uma variedade de outras fontes, além de extrairmos dados de estudos reais para desenvolver explicações e criar exercícios que demonstrem os muitos usos da estatística aplicada em administração e economia. Acreditamos que o emprego de dados reais ajuda a gerar no aluno maior interesse no material e lhe possibilita aprender tanto sobre a metodologia estatística como sobre sua aplicação. Esta edição contém mais de 300 exemplos e exercícios baseados em dados reais.

Novos Estudos de Caso

Adicionamos quatro novos estudos de caso nesta edição, chegando ao total de 21. Os estudos de caso aparecem nos capítulos sobre estatística descritiva, distribuição de probabilidade e regressão. Esses estudos de caso darão aos alunos a oportunidade de analisar conjuntos de dados relativamente maiores e preparar relatórios gerenciais baseados nos resultados da análise.

Novas Seções Estatística na Prática

Cada capítulo começa com uma seção chamada "Estatística na Prática", que descreve uma aplicação sobre a metodologia estatística a ser abordada. Os exemplos dessa seção foram fornecidos por profissionais de empresas como Colgate-Palmolive, Citibank, Procter & Gamble, Monsanto e outras. Esta edição inclui dois novos exemplos em "Estatística na Prática": Food Lion (Capítulo 8) e John Morrell & Company (Capítulo 9).

Materiais Novos para Planilhas de Cálculo do Minitab e do Microsoft® Excel

Apêndices das planilhas do Minitab e do Excel aparecem no final da maioria dos capítulos. Cada apêndice oferece instruções passo-a-passo que tornam o uso do Minitab ou do Excel mais fácil para os alunos, de modo que possam realizar as análises estatísticas apresentadas no capítulo. Todos os apêndices foram atualizados para as últimas versões do Minitab e do Excel. Sete apêndices novos e/ou revisados foram adicionados no final dos capítulos sobre estatística descritiva, estimação por intervalo, teste de hipóteses e regressão.

Características e Pedagogia

Continuamos com muitas das características introduzidas em edições anteriores. Algumas das mais importantes são destacadas a seguir.

Exercícios de Métodos e Exercícios de Aplicações

Os exercícios ao final de cada seção estão divididos em duas partes: Métodos e Aplicações. Os exercícios de Métodos exigem que os alunos usem as fórmulas e façam os cálculos necessários. Os exercícios de Aplicações requerem que os alunos usem o material do capítulo em situações reais. Desse modo, eles focalizam primeiro as generalidades computacionais e depois se dedicam às sutilezas da aplicação e interpretação estatística.

Exercícios de Autoteste

Certos exercícios são identificados como Autoteste. Soluções completas para tais exercícios são oferecidas no Apêndice D no final do livro. Os alunos podem tentar resolvê-los e imediatamente verificar as respostas para avaliar sua compreensão dos conceitos apresentados no capítulo.

Notas e Comentários

No final de muitas seções, acrescentamos notas e comentários que foram planejados para fornecer ao aluno critérios adicionais sobre a metodologia estatística e sua aplicação. A seção "Notas e Comentários" inclui avisos sobre as limitações da metodologia, recomendações para as aplicações, descrições suscintas de considerações técnicas adicionais e outros assuntos.

Conjuntos de Dados Acompanham o Livro

Aproximadamente 160 conjuntos de dados estão disponíveis para alunos e professores que utilizam esta obra, tanto em formato Minitab como em formato Excel. Ao longo do texto são usados logotipos denominados Arquivos da Internet para identificar este material, que se encontra no site:
www.thomsonlearning.com.br/estatapl.htm

Agradecimentos

Gostaríamos de agradecer o trabalho dos revisores da edição norte-americana, que forneceram comentários e sugestões de como continuar a melhorar nosso texto.

Michael Broida, Miami University of Ohio
Robert Cochran, University of Wyoming
Doug Dotterweich, East Tennessee State University
Dwight Goehring, California State University–Monterey Bay
Stephen Grubagh, Bentley University
David Keswick, University of Michigan–Flint
Jennifer Kohn, Montclair State University
Rosa Lemel, Kean University
Doug Morris, University of New Hampshire
Kevin Murphy, Oakland University
William Pan, University of New Haven
Carl Poch, Northern Illinois University
Ranga Ramasesh, Texas Christian University
Carolyn Rochelle, East Tennessee State University
Edwin Shapiro, University of San Francisco
Harvey Singer, George Mason University

Minghe Sun, University of Texas–San Antonio
Bill Swank, George Mason University
Geetha Vaidyanathan, University of North Carolina–Greensboro
James Wright, Green Mountain College
Elaine Zanutto, University of Pennsylvania

Continuamos em débito com nossos diversos colegas e amigos por seus comentários e sugestões úteis para o desenvolvimento desta edição e das anteriores de nosso livro. Entre eles citamos:

Ali Arshad, College of Santa Fe
Timothy M. Bergquist, Northwest Christian College
Habtu Braha, Coppin State College
Robert Cochran, University of Wyoming
Dale DeBoer, University of Colorado–Colorado Springs
Raj Devasagayam, St. Norbert College
Nader D. Ebrahimi, University of New Mexico
H. Robert Gadd, Southern Adventist University
Karen Gutermuth, Virginia Military Institute
Md. Mahbubul Kabir, Lyon College
Gary Nelson, Central Community College–Columbus Campus
Elaine Parks, Laramie County Community College
Charles J. Reichert, University of Wisconsin–Superior
Larry Scheuermann, University of Louisiana, Lafayette
Carlton Scott, University of California–Irvine
Alan D. Smith, Robert Morris College
Stephen L. S. Smith, Gordon College
Wibawa A. Sutanto, Prairie View A&M University
Bennie D. Waller, Francis Marion University
Yan Yu, University of Cincinnati
Zhiwei Zhu, University of Louisiana at Lafayette
Charles Zimmerman, Robert Morris College

Agradecemos especialmente aos nossos amigos da administração e indústria que forneceram as características da seção "Estatística na Prática". Nós os identificamos individualmente em cada um dos artigos. Finalmente, somos gratos ao nosso editor sênior de compras, Charles McCormick Jr., nossa editora sênior de produção, Deanna Quinn, nosso gerente sênior de marketing, Larry Qualls, e outros da Thomson Business and Professional Publishing por suas recomendações editoriais e apoio durante a preparação do texto.

David R. Anderson
Dennis J. Sweeney
Thomas A. Williams

Sobre os Autores

David R. Anderson é professor de análise quantitativa na Faculdade de Administração de Negócios da University of Cincinnati. Nascido em Grand Forks, Dakota do Norte, obteve os diplomas de bacharel, mestre e doutor na Purdue University. Exerceu os cargos de chefe do Departamento de Análise Quantitativa e Gerenciamento de Operações e de diretor-adjunto da Faculdade de Administração de Negócios, além de ter sido o coordenador do primeiro programa executivo dessa faculdade.

Na University of Cincinnati, o professor Anderson ensina estatística introdutória para os alunos de Administração, bem como ministra cursos de pós-graduação em Análise de Regressão, Análise Multivariada e Ciência da Administração. Ele também ministra cursos de estatística no Ministério do Trabalho, em Washington, D.C., e tem recebido honrarias como indicações e prêmios de excelência no ensino e excelência em serviços a organizações estudantis.

Anderson é co-autor de dez livros nas áreas de Estatística, Ciência da Administração, Programação Linear e Administração da Produção e Gerenciamento de Operações. É um ativo consultor no campo de amostragem e métodos estatísticos.

Dennis J. Sweeney é professor de análise quantitativa e fundador do Centro para a Melhoria da Produtividade na University of Cincinnati. Nascido em Des Moines, Iowa, obteve o diploma de bacharel na Drake University e os de mestre e doutor na Indiana University, onde foi membro do NDEA. Durante 1978 a 1979, integrou o grupo de Ciência da Administração da Procter & Gamble; de 1981 a 1982 foi professor visitante na Duke University. Exerceu os cargos de chefe do Departamento de Análise Quantitativa e diretor associado da Faculdade de Administração de Negócios da University of Cincinnati.

Sweeney publicou mais de 30 artigos e monografias na área de Ciência da Administração e Estatística. A National Science Foundation, IBM, Procter & Gamble, Federated Department Stores, Kroger e Cincinnati Gas & Electric têm patrocinado suas pesquisas, publicadas nas revistas *Management Science, Operations Research, Mathematical Programming* e *Decision Science,* entre outras.

É co-autor de oito livros nas áreas de Estatística, Ciência de Gerenciamento, Programação Linear e Administração da Produção e de Operações.

Thomas A. Williams é professor de ciência da administração na Faculdade de Administração do Rochester Institute of Technology (RIT). Nascido em Elmira, Nova York, obteve o diploma de bacharel na Clarkson University. Fez pós-graduação no Rensselaer Polytechnic Institute, onde obteve os diplomas de mestre e doutor.

Antes de lecionar na Faculdade de Administração do RIT, Williams foi, durante sete anos, professor da Faculdade de Administração de Negócios da University of Cincinnati, onde desenvolveu e coordenou o

programa de Sistemas de Informação. No RIT, foi o primeiro presidente do Departamento de Ciências da Decisão. Ministra cursos de Ciência da Administração e de Estatística, bem como cursos de graduação em Análise de Regressão e de Decisão.

É co-autor de 11 livros nas áreas de Ciência da Administração, Estatística, Administração da Produção e de Operações e Matemática. É consultor de numerosas empresas da *Fortune 500* e tem trabalhado em projetos que vão do uso da análise de dados ao desenvolvimento de modelos de regressão de grande escala.

Sumário

Capítulo 1 Dados e a Estatística

Estatística na Prática: *Business Week* 1
1.1 Aplicações em Administração e Economia 2
Contabilidade 3
Finanças 3
Marketing 3
Produção 3
Economia 3
1.2 Dados 4
Elementos, Variáveis e Observações 4
Escalas de Medição 5
Dados Qualitativos e Quantitativos 6
Dados de Seção Transversal e de Série Histórica 6
1.3 As Fontes de Dados 7
Fontes Existentes 7
Estudos Estatísticos 8
Erros na Obtenção de Dados 10
1.4 Estatística Descritiva 10
1.5 Inferência Estatística 12
1.6 Computadores e a Análise Estatística 13
Resumo 13
Glossário 14
Exercícios 15

Capítulo 2 Estatística Descritiva: Métodos Tabulares e Métodos Gráficos

Estatística na Prática: A Companhia Colgate-Palmolive 21
2.1 Sintetizando os Dados Qualitativos 23
A Distribuição de Freqüência 23
As Distribuições de Freqüência Relativa e de Freqüência Percentual 24
Gráficos em Barras e em Setores ("Pizza") 24
2.2 Sintetizando os Dados Quantitativos 28
A Distribuição de Freqüência 28
As Distribuições de Freqüência Relativa e de Freqüência Percentual 30
Gráficos de Dispersão Unidimensional 31
Histograma 31
Distribuições Cumulativas 32
Ogivas 34
2.3 Análise Exploratória dos Dados: A Apresentação de Ramo-e-Folha 38
2.4 Tabulações Cruzadas e Diagramas de Dispersão 43
Tabulação Cruzada 43
O Paradoxo de Simpson 45
Diagramas de Dispersão e Linha de Tendência 46
Resumo 52
Glossário 53
Fórmulas-Chave 53
Exercícios Suplementares 54
Estudo de Caso – Pelican Stores 59
Apêndice 2.1 – O Uso do Minitab para Apresentações Tabulares e Gráficas ... 60
Apêndice 2.2 – O Uso do Excel para Apresentações Tabulares e Gráficas..... 62

Capítulo 3 Estatística Descritiva: Medidas Numéricas

Estatística na Prática: Small Fry Design 71
3.1 Medidas de Posição 72
Média 72
Mediana 74
Moda 74
Percentis 75
Quartis 76
3.2 Medidas de Variabilidade 81
Amplitude 81
Amplitude Interquartil 82
Variância 82
Desvio Padrão 84
Coeficiente de Variação 84
3.3 Medidas da Forma da Distribuição, da Posição Relativa e Detecção de Pontos Fora da Curva 87
Forma da Distribuição 87
Contagens-*z* 88

Teorema de Chebyshev .89
Regra Empírica .90
Detecção de Pontos Fora da Curva .90
3.4 Análise Exploratória de Dados .94
Regra de Cinco Itens .94
Desenhos Esquemáticos (*Box Plots*) .94
3.5 Medidas de Associação entre Duas Variáveis .98
Covariância .98
Interpretação da Covariância .99
Coeficiente de Correlação .102
Interpretação do Coeficiente de Correlação .102
3.6 Média Ponderada e o Trabalho com Dados Agrupados .106
Média Ponderada .106
Dados Agrupados .107
Resumo .111
Glossário .111
Fórmulas-Chave .112
Exercícios Suplementares .114
Estudo de Caso 1 – Pelican Stores .118
Estudo de Caso 2 – National Health Care Association .119
Estudo de Caso 3 – Escolas de Administração da Região Ásia-Pacífico .120
Apêndice 3.1 – Estatística Descritiva com o Minitab .122
Apêndice 3.2 – Estatística Descritiva com o Excel .124

Capítulo 4 Introdução à Probabilidade

Estatística na Prática: Morton International .129
4.1 Experimentos, Regras de Contagem e Atribuindo Probabilidades .130
Regras de Contagem, Combinações e Permutações .131
Atribuindo Probabilidades .135
Probabilidades do Projeto da KP&L .136
4.2 Eventos e Suas Probabilidades .139
4.3 Algumas Relações Básicas de Probabilidade .143
Complemento de um Evento .143
Lei da Adição .143
4.4 Probabilidade Condicional .148
Eventos Independentes .151
Lei da Multiplicação .151
4.5 Teorema de Bayes .155
A Abordagem Tabular .158
Resumo .160
Glossário .160
Fórmulas-Chave .161
Exercícios Suplementares .162
Estudo de Caso – Os Juízes do Condado de Hamilton .166

Capítulo 5 Distribuições Discretas de Probabilidade

Estatística na Prática: Citibank 169
5.1 Variáveis Aleatórias 170
Variáveis Aleatórias Discretas 170
Variáveis Aleatórias Contínuas 171
5.2 Distribuições Discretas de Probabilidade 173
5.3 Valor Esperado e Variância 177
Valor Esperado 177
Variância 178
5.4 Distribuição de Probabilidade Binomial 181
Um Experimento Binomial 182
O Problema da Loja de Roupas do Martin 183
Usando Tabelas de Probabilidades Binomiais 187
Valor Esperado e Variância da Distribuição Binomial 188
5.5 Distribuição de Poisson 191
Um Exemplo Envolvendo Intervalos de Tempo 192
Um Exemplo Envolvendo Intervalos de Comprimento ou de Distância 193
5.6 Distribuição de Probabilidade Hipergeométrica 195
Resumo 198
Glossário 198
Fórmulas-Chave 199
Exercícios Suplementares 200
Apêndice 5.1 – Distribuições Discretas de Probabilidade com o Minitab 202
Apêndice 5.2 – Distribuições Discretas de Probabilidade com o Excel 202

Capítulo 6 Distribuições Contínuas de Probabilidade

Estatística na Prática: Procter & Gamble 205
6.1 Distribuição Uniforme de Probabilidade 207
A Área como uma Medida de Probabilidade 207
6.2 Distribuição Normal de Probabilidade 210
Curva Normal 211
Distribuição Normal Padrão de Probabilidade 212
Como Calcular Probabilidades de Qualquer Distribuição Normal 218
O Problema da Grear Tire Company 218
6.3 Aproximação Normal às Probabilidades Binomiais 223
6.4 Distribuição Exponencial de Probabilidade 226
Como Calcular Probabilidades da Distribuição Exponencial 226
Relações entre a Distribuição de Poisson e a Distribuição Exponencial 228
Resumo 230
Glossário 230
Fórmulas-Chave 230
Exercícios Suplementares 231
Estudo de Caso – Specialty Toys 233
Apêndice 6.1 – Distribuições Contínuas de Probabilidade com o Minitab 234
Apêndice 6.2 – Distribuições Contínuas de Probabilidade com o Excel 235

Capítulo 7 Amostragens e Distribuições Amostrais

Estatística na Prática: MeadWestvaco Corporation 237
7.1 Problema de Amostragem da Electronics Associates 239
7.2 Amostragem Aleatória Simples 239
Amostragem de Populações Finitas 240
Amostragem de Populações Infinitas 241
7.3 Estimação por Ponto 244
7.4 Introdução às Distribuições Amostrais 246
7.5 Distribuição Amostral de $\bar{x}$ 249
Valor Esperado de $\bar{x}$ 249
Desvio Padrão de $\bar{x}$ 249
Forma da Distribuição Amostral de $\bar{x}$ 250
Distribuição Amostral de $\bar{x}$ para o Problema da EAI 252
Valor Prático da Distribuição Amostral de $\bar{x}$ 252
Relação entre o Tamanho da Amostra e a Distribuição Amostral de $\bar{x}$ 253
7.6 Distribuição Amostral de $\bar{p}$ 257
Valor Esperado de $\bar{p}$ 258
Desvio Padrão de $\bar{p}$ 258
Forma da Distribuição Amostral de $\bar{p}$ 259
Valor Prático da Distribuição Amostral de $\bar{p}$ 259
7.7 Métodos de Amostragem 262
Amostragem Aleatória Estratificada 262
Amostragem por Conglomerados 263
Amostragem Sistemática 263
Amostragem de Conveniência 264
Amostragem de Julgamento 264
Resumo 265
Glossário 265
Fórmulas-Chave 266
Exercícios Suplementares 266
Apêndice 7.1 – Amostragem Aleatória com o Minitab 268
Apêndice 7.2 – Amostragem Aleatória com o Excel 268

Capítulo 8 Estimação por Intervalo

Estatística na Prática: Food Lion 271
8.1 Média da População: σ Conhecido 272
Margem de Erro e a Estimação por Intervalo 273
Conselho Prático 276
8.2 Média da População: σ Desconhecido 278
Margem de Erro e a Estimação por Intervalo 279
Conselho Prático 282
Como Usar uma Amostra Pequena 282
Resumo dos Procedimentos de Estimação 284
8.3 Como Determinar o Tamanho da Amostra 287

8.4 Proporção da População 290
Como Determinar o Tamanho da Amostra 291
Resumo 295
Glossário 296
Fórmulas-Chave 296
Exercícios Suplementares 297
Estudo de Caso 1 – Bock Investment Services 300
Estudo de Caso 2 – Gulf Real Estate Properties 300
Estudo de Caso 3 – Metropolitan Research, Inc. 303
Apêndice 8.1 – Estimação por Intervalo com o Minitab 303
Apêndice 8.2 – Estimação por Intervalo com o Excel 305

Capítulo 9 Testes de Hipóteses

Estatística na Prática: John Morrell & Company 309
9.1 Como Desenvolver as Hipóteses Nula e Alternativa 310
Como Testar Hipóteses de Pesquisa 310
Como Testar a Validade de uma Afirmação 311
Como Testar em Situações de Tomada de Decisão 311
Resumo das Formas das Hipóteses Nula e Alternativa 311
9.2 Erros do Tipo I e do Tipo II 313
9.3 Média da População: σ Conhecido 315
Teste Unicaudal 315
Teste Bicaudal 320
Resumo e Conselho Prático 323
Relação entre a Estimação por Intervalo e o Teste de Hipóteses 324
9.4 Média da População: σ Desconhecido 328
Teste Unicaudal 329
Teste Bicaudal 330
Resumo e Conselho Prático 331
9.5 Proporção da População 335
Resumo 337
Resumo 340
Glossário 340
Fórmulas-Chave 341
Exercícios Suplementares 341
Estudo de Caso 1 – Quality Associates, Inc. 343
Estudo de Caso 2 – Estudo do Desemprego 345
Apêndice 9.1 – Testes de Hipóteses com o Minitab 345
Apêndice 9.2 – Testes de Hipóteses com o Excel 347

Capítulo 10 Comparações Envolvendo Médias

Estatística na Prática: Fisons Corporation 353
10.1 Inferências sobre a Diferença entre as Médias
de Duas Populações: σ_1 e σ_2 Conhecidos 354

Estimação por Intervalo de $\mu_1 - \mu_2$ .354
Testes de Hipóteses sobre $\mu_1 - \mu_2$ .356
Conselho Prático .358
10.2 Inferências sobre a Diferença entre as Médias de Duas Populações: σ_1 e σ_2 Desconhecidos .360
Estimação por Intervalo de $\mu_1 - \mu_2$ .360
Testes de Hipóteses sobre $\mu_1 - \mu_2$ .361
Conselho Prático .363
10.3 Inferências sobre a Diferença entre as Médias de Duas Populações: Amostras Relacionadas (ou Dependentes) .368
10.4 Introdução à Análise de Variância .372
Hipóteses sobre a Análise de Variância .374
Visão Conceitual .374
10.5 Análise de Variância: Como Testar a Igualdade de k Médias da População .376
Estimativa da Variância Populacional entre Tratamentos .377
Estimativa da Variância Populacional dentro de Tratamentos .378
Comparando as Estimativas de Variância: o Teste F .378
A Tabela ANOVA .381
Resultados de Computador para a Análise de Variância .381
Resumo .385
Glossário .386
Fórmulas-Chave .386
Exercícios Suplementares .388
Estudo de Caso 1 – Par, Inc. .391
Estudo de Caso 2 – Wentworth Medical Center .392
Estudo de Caso 3 – Remuneração de Profissionais de ID .393
Apêndice 10.1 – Inferências sobre Duas Populações com o Minitab .394
Apêndice 10.2 – Inferências sobre Duas Populações com o Excel .395
Apêndice 10.3 – Análise de Variância com o Minitab .396
Apêndice 10.4 – Análise de Variância com o Excel .396

Capítulo 11 Comparações Envolvendo Proporções e Teste de Independência

Estatística na Prática: United Way .399
11.1 Inferências sobre a Diferença entre as Proporções de Duas Populações .400
Estimação por Intervalo de $p_1 - p_2$ .400
Testes de Hipóteses sobre $p_1 - p_2$ .402
11.2 Testes de Hipóteses para Proporções de uma População Multinomial .406
11.3 Teste de Independência .411
Resumo .417
Glossário .418
Fórmulas-Chave .418
Exercícios Suplementares .418
Estudo de Caso – Programa Bipartidário de Reforma .422

Apêndice 11.1 – Inferências sobre Duas Proporções Populacionais com o Minitab .423
Apêndice 11.2 – Testes de Eficiência de Ajuste e de Independência com o Minitab .424
Apêndice 11.3 – Testes de Eficiência de Ajuste e de Independência com o Excel .424

Capítulo 12 Regressão Linear Simples

Estatística na Prática: Alliance Data Systems .427
12.1 Modelo de Regressão Linear Simples .428
Modelo de Regressão e Equação de Regressão .429
Equação de Regressão Estimada .429
12.2 Método dos Mínimos Quadrados .431
12.3 Coeficiente de Determinação .440
Coeficiente de Correlação .443
12.4 Suposições do Modelo .447
12.5 Teste de Significância .448
Estimativa de σ^2 .448
Teste t .449
Intervalo de Confiança de β_1 .450
Teste F .451
Alguns Cuidados com a Interpretação dos Testes de Significância .452
12.6 Usando a Equação de Regressão Estimada para Estimação e Previsão .456
Estimação por Ponto .456
Estimação por Intervalo .456
Intervalo de Confiança do Valor Médio de y .456
Intervalo de Previsão para um Valor Individual de y .458
12.7 Solução Computadorizada .462
12.8 Análise Residual: Validando Suposições do Modelo .466
Plotagem Residual em Relação a x .467
Plotagem Residual em Relação a $\hat{y}$ .468
Resumo .471
Glossário .471
Fórmulas-Chave .472
Exercícios Suplementares .473
Estudo de Caso 1 – Gastos e Desempenho Escolar .478
Estudo de Caso 2 – U.S. Department of Transportation .480
Estudo de Caso 3 – Doações de Ex-Alunos .480
Estudo de Caso 4 – Valores do Times de Beisebol da Major League .482
Apêndice 12.1 – Análise de Regressão com o Minitab .483
Apêndice 12.2 – Análise de Regressão com o Excel .484

Capítulo 13 Regressão Múltipla

Estatística na Prática: International Paper .487
13.1 Modelo de Regressão Múltipla .488

Modelo de Regressão e Equação de Regressão .488
Equação de Regressão Múltipla Estimada .489
13.2 Método dos Mínimos Quadrados .489
Exemplo: Butler Trucking Company .490
Nota sobre a Interpretação de Coeficientes .492
13.3 Coeficiente de Determinação Múltiplo .497
13.4 Suposições do Modelo .500
13.5 Teste de Significância .501
Teste *F* .502
Teste *t* .504
Multicolinearidade .504
13.6 Usando a Equação de Regressão Estimada para Estimação e Previsão .507
13.7 Variáveis Qualitativas Independentes .509
Exemplo: Johnson Filtration, Inc. .509
Interpretando os Parâmetros .511
Variáveis Qualitativas mais Complexas .513
Resumo .517
Glossário .517
Fórmulas-Chave .518
Exercícios Suplementares .519
Estudo de Caso 1 – Consumer Research, Inc. .523
Estudo de Caso 2 – Previsão das Pontuações no Exame de Proficiência Escolar .524
Estudo de Caso 3 – Doações de Ex-Alunos .525
Apêndice 13.1 – Regressão Múltipla com o Minitab .525
Apêndice 13.2 – Regressão Múltipla com o Excel .526

Apêndices

Apêndice A Referências e Bibliografia .529
Apêndice B Tabelas .531
Apêndice C Notação de Somatório .555
Apêndice D Soluções dos Autotestes e Respostas dos Exercícios Pares .559

Índice Remissivo .589

CAPÍTULO 1

Dados e a Estatística

ESTATÍSTICA NA PRÁTICA

*BUSINESS WEEK**
Nova York, NY

Com uma circulação global de mais de um milhão de exemplares, a *Business Week* é a revista de negócios mais lida em todo o mundo. Mais de 200 repórteres exclusivos e editores em 26 agências internacionais publicam uma série de artigos que interessam à comunidade empresarial e econômica. Além dos artigos especiais sobre temas da atualidade, a revista contém seções regulares sobre negócios internacionais, análise econômica, processamento de informação e ciência e tecnologia. As informações apresentadas nos artigos e nas seções regulares ajudam o leitor a manter-se atualizado sobre os acontecimentos e a avaliar o impacto desses acontecimentos sobre a economia e os negócios.

A maioria das edições da *Business Week* fornece uma reportagem mais aprofundada sobre um assunto de interesse atual. Freqüentemente essas reportagens contêm fatos e resumos estatísticos que ajudam o leitor a entender a informação empresarial ou econômica. Por exemplo, a edição de 11 de novembro de 2003 trouxe uma reportagem sobre o novo impulso das comunicações sem fio; a edição de 15 de dezembro de 2003 publicou sobre os melhores produtos de 2003; a edição de 12 de janeiro de 2004 descreveu o panorama econômico para 2004, conforme a visão da indústria; e a edição de 26 de janeiro de 2004 continha informações sobre os melhores fundos mútuos para o ano seguinte. Além disso, a seção semanal *Business Week Investor* apresenta dados estatísticos sobre a economia, incluindo índices de produção, preços de ações, fundos mútuos e taxas de juros.

A *Business Week* também usa a estatística e informações estatísticas para gerenciar seu próprio negócio. Por exemplo, uma pesquisa anual feita com os assinantes ajuda a empresa a conhecer aspectos demográficos relativos a eles, seus hábitos de leitura, probabilidade de compras, estilos de vida e assim por diante. Os gerentes da *Business Week* utilizam os sumários estatísticos dessa pesquisa para oferecer melhores serviços aos assinantes e aos anunciantes.

* Os autores agradecem a Charlene Trentham, gerente de Pesquisas da *Business Week*, por fornecer esta "Estatística na Prática".

Uma pesquisa recente com os assinantes norte-americanos indicou que 90% dos assinantes da *Business Week* têm computadores em casa e que 64% articulam a compra de um computador no trabalho. Esse tipo de estatística alerta os gerentes da *Business Week* quanto ao interesse do assinante em artigos sobre novos desenvolvimentos na área da informática. Os resultados da pesquisa também são colocados à disposição de potenciais assinantes. A elevada porcentagem de assinantes que usam computadores em casa e dos que articulam a compra de computadores no trabalho seria um incentivo para os fabricantes de computadores pensarem em anunciar na *Business Week*.

Neste capítulo, discutiremos os tipos de dados disponíveis para análise estatística e descreveremos como eles são obtidos. Apresentaremos a estatística descritiva e a inferência estatística como meios de se converter dados em informações estatísticas significativas e de fácil interpretação.

Vemos com freqüência os seguintes tipos de afirmação em artigos de jornais e de revistas:

- Uma pesquisa realizada pela Jupiter Media descobriu que 31% dos homens adultos passam dez ou mais horas por semana assistindo à televisão. Em relação às mulheres adultas, o resultado foi 26% (*The Wall Street Journal*, 26 de janeiro de 2004).
- A General Motors, líder em descontos para carros de passeio, apresentou uma média de US$ 4.300 de incentivo financeiro para a compra de veículos durante o ano de 2003 (*USA Today*, 23 de janeiro de 2004).
- Mais de 40% dos gerentes da Marriott International iniciaram a carreira como funcionários de baixo escalão (*Fortune*, 20 de janeiro de 2003).
- Os empregos no setor de administração e finanças tiveram uma média de US$ 49.712 quanto ao salário anual para 2003 (*The World Almanac*, 2004).
- Os empregadores planejavam contratar 12,7% mais pessoas com graduação universitária em 2004 do que em 2003 (Collegiate Employment Research Institute, Michigan State University, fevereiro de 2004).
- A equipe dos New York Yankees tem a folha de pagamento mais cara da principal liga de beisebol. Em 2003, a folha de pagamento da equipe foi de US$ 152.749.814, com uma média de US$ 4.575 mil por jogador (USA Today, 1º de setembro de 2003).
- A Média Industrial Dow Jones (Dow Jones Industrial Average) fechou em 10.358 em 31 de março de 2004 (*The Wall Street Journal*, 1º de abril de 2004).

Os fatos numéricos contidos nas afirmações dadas (31%, 26%, US$ 4.300, 40%, US$ 49.712, 12,7%, US$ 152.749.814, 4.575 mil e 10.358, denominam-se estatísticas. Desse modo, no uso diário o termo **estatística** refere-se a fatos numéricos. Entretanto, a área ou o tema da estatística envolve muito mais do que fatos numéricos. Em um sentido amplo, estatística é a arte e a ciência de coletar, analisar, apresentar e interpretar dados. Especialmente na área da administração e economia, as informações obtidas por meio de coleta, análise, apresentação e interpretação dos dados proporcionam aos gerentes e tomadores de decisões uma melhor compreensão do ambiente empresarial e econômico e, assim, capacita-os a tomar decisões mais fundamentadas e de melhor qualidade. Neste livro, enfatizamos o uso da estatística para tomar decisões nas áreas de administração e economia.

O Capítulo 1 inicia-se com algumas ilustrações da aplicação da estatística no setor de administração e economia. Na Seção 1.2, definimos o termo *dados* e introduzimos o conceito de conjunto de dados. Essa seção também apresenta termos-chave, tais como *variáveis e observações*, discute a diferença entre dados quantitativos e qualitativos e ilustra o uso de dados transversais e de séries históricas. A Seção 1.3 discute como é possível obter dados de fontes existentes ou por intermédio de pesquisa e estudos experimentais idealizados para obter novos dados. O importante papel que a Internet agora desempenha na obtenção de dados também é realçado. A utilização de dados para desenvolver estatística descritiva e para se fazer inferências estatísticas será descrita nas Seções 1.4 e 1.5.

1.1 APLICAÇÕES EM ADMINISTRAÇÃO E ECONOMIA

No moderno ambiente administrativo e econômico global, qualquer pessoa pode ter acesso a uma enorme quantidade de informações estatísticas. Os gerentes e tomadores de decisão mais bem-sucedidos são aque-

les capazes de entender a informação e usá-la eficazmente. Nesta seção, apresentamos exemplos que ilustram algumas utilizações da estatística nas áreas da administração e economia.

Contabilidade

Empresas públicas de contabilidade utilizam procedimentos de amostragem estatística ao realizarem auditorias para seus clientes. Por exemplo, suponha que uma firma de contabilidade queira determinar se o valor das contas a receber indicado na folha de balancete de um cliente representa fielmente o valor real das contas a receber. Geralmente o grande número de contas a receber individuais torna a revisão e validação de cada conta algo demasiadamente demorado e dispendioso. A prática comum nessas situações é a equipe de auditores selecionar um subconjunto das contas, denominado amostra. Depois de revisar a exatidão das contas amostradas, os auditores concluem se o valor das contas a receber apresentado na folha de balancete do cliente é aceitável.

Finanças

Os analistas financeiros usam uma série de informações estatísticas para orientar suas recomendações de investimentos. No caso dos títulos financeiros, os analistas revisam uma série de dados financeiros que incluem os índices de preço/ganhos ou lucros e a rentabilidade em dividendos. Comparando a informação correspondente a um título individual com as informações sobre a média do mercado de ações, o analista financeiro pode concluir se um título individual está valorizado ou desvalorizado.

Por exemplo, a revista *Barron's* (6 de janeiro de 2003) publicou que a média dos índices de preço/ganhos ou lucros dos 30 títulos da Média Industrial Dow Jones era de 22,36. A General Electric apresentava um índice de preço/ganhos ou lucros igual a 16. Neste caso, a informação estatística sobre os índices de preço/ganhos ou lucros indicava um preço comparativamente menor para os ganhos ou lucros da General Electric em comparação aos títulos da Dow Jones. Portanto, um analista financeiro poderia concluir que os títulos da General Electric estavam desvalorizados. Esta e outras informações sobre a General Electric ajudariam o analista a recomendar a compra, venda ou manutenção dos títulos.

Marketing

Scanners eletrônicos utilizados nas caixas registradoras das lojas de venda a varejo coletam dados que são usados em uma série de aplicações de pesquisa de marketing. Por exemplo, fornecedores de dados como a ACNielsen e a Information Resources Inc. compram dados colhidos por *scanners* eletrônicos localizados em pontos-de-venda de mercearias, processam esses dados e depois vendem seus sumários estatísticos a empresas de manufatura. Empresas manufatureiras gastam centenas de milhares de dólares por categoria de produto para obter esse tipo de informação. A indústria também compra dados e sumários estatísticos a respeito de atividades promocionais, como a fixação de preços especiais e o uso de exibições em vídeo nas lojas. Gerentes de marca podem revisar os dados estatísticos dos *scanners* e os dados estatísticos da atividade promocional para obter um entendimento melhor da relação entre as atividades promocionais e as vendas. Esse tipo de análise muitas vezes é útil para estabelecer as futuras estratégias de marketing para os vários produtos.

Produção

A atual ênfase na qualidade torna o controle da qualidade uma importante aplicação da estatística na área de produção. Usa-se uma série de mapas estatísticos de controle da qualidade para monitorar o resultado (*output*) de um processo de produção. Em especial, pode-se usar um gráfico de barras para monitorar a média do produto. Suponha, por exemplo, que uma máquina preencha recipientes com 340 ml de determinado refrigerante. Periodicamente, um funcionário do setor de produção seleciona uma amostra dos recipientes e calcula a quantidade média de refrigerante em mililitros. Essa média, ou valor de barra, é traçada em um gráfico de barras. Um valor acima do limite máximo de controle no gráfico mostra que o recipiente tem um volume de refrigerante maior que o especificado, e um valor abaixo do limite mínimo de controle no gráfico mostra que o recipiente tem um volume menor do que o especificado. O processo é chamado "sob controle" e pode prosseguir contanto que os valores de barras traçados se situem entre os limites de controle máximo e mínimo indicados no gráfico. Adequadamente interpretado, um gráfico de barras pode ajudar a estabelecer quando há a necessidade de ajustes para corrigir o processo de produção.

Economia

Os economistas freqüentemente fornecem previsões sobre o futuro da economia ou algum aspecto dela. Eles usam uma série de informações estatísticas para fazer essas previsões. Por exemplo, ao preverem as taxas de

são significativas. Por exemplo, o estudante 1 pontuou 1.120 – 1.050 = 70 pontos a mais que o estudante 2, ao passo que o estudante 2 pontuou 1.050 – 970 = 80 pontos a mais que o estudante 3.

A escala de medição de uma variável é uma **escala de proporção** se os dados tiverem todas as propriedades de dados de intervalo e a proporção de dois valores for significativa. Variáveis como distância, altura, peso e tempo usam como medição a escala de proporção. Essa escala exige que um valor zero seja incluído para indicar que não existe nada para a variável no ponto zero. Por exemplo, considere o custo de um automóvel. Um valor zero para o custo indicaria que o automóvel não tem nenhum custo e é grátis. Além disso, se compararmos o custo de US$ 30 mil para um automóvel com o custo de US$ 15 mil para um segundo automóvel, a propriedade da razão mostra que o primeiro automóvel é US$ 30 mil/ US$ 15 mil = 2 vezes (ou o dobro) o custo do segundo automóvel.

Dados Qualitativos e Quantitativos

Os dados podem ser adicionalmente classificados como qualitativos ou quantitativos. Os **dados qualitativos** incluem rótulos ou nomes usados para identificar um atributo de cada elemento. Os dados qualitativos utilizam a escala de medição nominal ou a ordinal e podem ser não-numéricos ou numéricos. **Dados quantitativos** requerem valores numéricos que indicam quantificação ou quantidade numérica. Dados quantitativos são obtidos usando-se ou a escala de medição intervalar ou a escala de proporção.

Os dados qualitativos freqüentemente são chamados dados categóricos.

Uma **variável qualitativa** é uma variável com dados qualitativos, e uma **variável quantitativa** é uma variável com dados quantitativos. A análise estatística apropriada de determinada variável depende de a variável ser qualitativa ou quantitativa. Se a variável for qualitativa, a análise estatística será bastante limitada. Podemos sintetizar os dados qualitativos contando o número de observações em cada categoria qualitativa ou calculando a proporção das observações em cada categoria qualitativa. Entretanto, mesmo quando os dados qualitativos usam um código numérico, operações aritméticas como a adição, subtração, multiplicação e divisão não produzem resultados significativos. A Seção 2.1 discute maneiras de sintetizar dados qualitativos.

O método estatístico apropriado para se fazer sumários de dados depende de os dados serem qualitativos ou quantitativos.

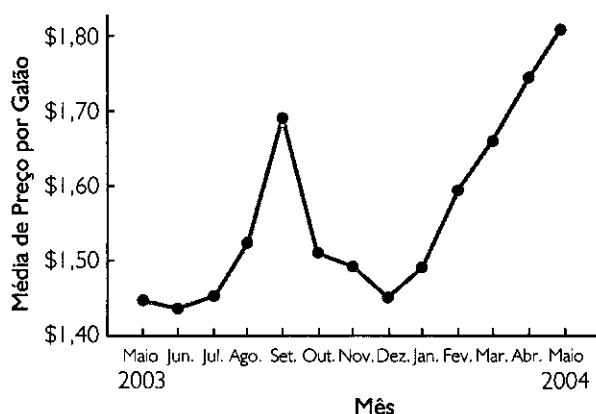
No entanto, operações aritméticas freqüentemente produzem resultados significativos para uma variável quantitativa. Por exemplo, em relação a uma variável quantitativa, os dados podem ser somados e depois divididos pelo número de observações para calcularmos o valor médio. Essa média geralmente é significativa e facilmente interpretada. Em geral, quando os dados são quantitativos há mais alternativas para a análise estatística. A Seção 2.2 e o Capítulo 3 apresentam maneiras de sintetizar dados quantitativos.

Dados de Seção Transversal e de Série Histórica

Para fins de análise estatística, é importante estabelecer a distinção entre dados de seção transversal e dados de série histórica. **Dados de seção transversal** são dados coletados no mesmo intervalo de tempo ou aproximadamente no mesmo intervalo de tempo. Os dados da Tabela 1.1 são transversais porque descrevem as cinco variáveis correspondentes aos 25 *shadow stocks* (títulos-fantasma) no mesmo intervalo de tempo. **Dados de série histórica** são dados coletados ao longo de diversos períodos. Por exemplo, a Figura 1.1 apresenta um gráfico da média de preço por galão de gasolina comum sem chumbo, nas cidades norte-americanas. O gráfico mostra uma abrupta elevação do preço médio por galão a partir de janeiro de 2004. Ao longo de um período de cinco meses, a média de preço por galão se elevou de US$ 1,49 para US$ 1,81. A maioria dos métodos estatísticos apresentados neste livro se aplica ao tipo de dados de seção transversal, não a dados de série histórica.

NOTAS E COMENTÁRIOS

1. Uma observação é o conjunto de medidas obtidas correspondentes a cada elemento de um conjunto de dados. Portanto, o número de observações é sempre igual ao número de elementos. O número de medidas obtidas correspondentes a cada elemento é igual ao número de variáveis. Portanto, o número total de itens de dados pode ser determinado multiplicando-se o número de observações pelo número de variáveis.
2. Os dados quantitativos podem ser discretos ou contínuos. Dados quantitativos que medem a quantidade são discretos. Dados quantitativos que medem a quantificação são contínuos, porque não ocorre nenhuma separação entre os valores de dados possíveis.

Figura 1.1 Preço médio por galão de gasolina comum sem chumbo, nas cidades norte-americanas

Fonte: U.S. Energy Information Administration, maio de 2004.

1.3 AS FONTES DE DADOS

Os dados podem ser obtidos de fontes existentes ou de pesquisas e estudos experimentais concebidos para esse fim.

Fontes Existentes

Em alguns casos, os dados necessários a uma aplicação em particular já existem. As empresas mantêm uma série de bancos de dados sobre seus empregados, clientes e operações empresariais. Dados sobre salários dos empregados, idade e experiência geralmente podem ser obtidos dos registros internos do departamento pessoal. Outros registros internos contêm dados sobre vendas, gastos com propaganda, custos de distribuição, níveis de estoque e quantidades de produção. A maioria das empresas também mantém dados detalhados a respeito de seus clientes. A Tabela 1.2 apresenta alguns dos dados que habitualmente estão disponíveis nos registros internos da empresa.

Organizações especializadas em coletar e manter dados disponibilizam uma quantidade substancial de dados empresariais e econômicos. As empresas têm acesso a essas fontes externas de dados por contratos de *leasing*[3], ou por meio de compra. A Dun & Bradstreet, a Bloomberg e a Dow Jones & Company são três firmas que oferecem amplos serviços de bancos de dados empresariais aos seus clientes.

Tabela 1.2 Exemplos de dados disponíveis nos registros internos das empresas

Fonte	Dados Tipicamente Disponíveis
Registros de funcionários	Nome, endereço, número do seguro social, número de dias de férias, número de dias dedicados a tratamento de saúde e bonificações.
Registros de produção	Número de peças ou produtos, quantidade produzida, custo de mão-de-obra e custo de matérias-primas.
Registros de estoques	Número de peças ou produtos, número de unidades disponíveis, nível de reencomenda, lote econômico de compra e programa de descontos.
Registros de vendas	Número do produto, volume de vendas, volume de vendas por região e volume de vendas por tipo de cliente.
Registros de crédito	Nome do cliente, endereço, número telefônico, limite de crédito e saldo de contas a receber.
Perfil do cliente	Idade, sexo, nível de renda, tamanho da família, endereço e preferências.

[3] NT: *Leasing* – Arrendamento mercantil.

A ACNielsen e a Information Resources, Inc. construíram negócios bem-sucedidos coletando e processando dados que são vendidos a empresas de publicidade e de manufatura.

Dados também se encontram disponíveis em uma série de associações industriais e organizações de interesse especial. A Travel Industry Association of America mantém informações relacionadas a viagens, por exemplo, o número de turistas e os gastos em viagens, organizados por Estado. Esses dados interessariam a firmas e a pessoas da indústria de viagens. O Graduate Management Admission Council mantém dados sobre notas de exames, características do estudante e programas de ensino de pós-graduação em administração. A maior parte dos dados desses tipos de fontes se encontra disponível a usuários habilitados a um pequeno custo.

A internet continua a se expandir como uma importante fonte de dados e de informações estatísticas. Quase todas as empresas mantêm *websites* que fornecem informações gerais sobre a empresa, bem como dados de vendas, número de empregados, número de produtos, preços dos produtos e especificações dos produtos. Além disso, agora, um grande número de empresas se especializa em tornar disponível informações pela rede. Em conseqüência, pode-se ter acesso a cotações de ações, preços de refeições em restaurantes, dados salariais e uma variedade quase infinita de informações.

Acredita-se que o maior estudo estatístico experimental já realizado tenha sido o experimento da vacina Salk, contra a poliomielite, promovido pelo Public Health Service (Estados Unidos) em 1954. Aproximadamente 2 milhões de crianças do primeiro, segundo e terceiro anos do ensino fundamental foram selecionadas em todo o território nacional.

Órgãos governamentais são outra fonte importante de dados existentes. Por exemplo, o U.S. Department of Labor (departamento do trabalho norte-americano) mantém dados consideráveis sobre os índices de emprego, índices salariais, tamanho da força trabalhista e afiliação sindical. A Tabela 1.3 relaciona os órgãos governamentais e alguns dos dados que eles oferecem. A maioria dos órgãos governamentais que coleta e processa dados também disponibiliza os resultados por meio de um site. Por exemplo, o U.S. Census Bureau (departamento do censo norte-americano) tem uma vasta quantidade de dados em seu endereço: www.census.gov. A Figura 1.2 exibe a página inicial do U.S. Census Bureau.

Estudos Estatísticos

Às vezes, os dados necessários a uma aplicação em particular não se encontram disponíveis por meio das fontes existentes. Nesses casos, freqüentemente os dados são obtidos pela realização de um estudo estatístico. Os estudos estatísticos podem ser classificados como *experimentais* ou *baseados na informação.*

Em um estudo experimental, identifica-se primeiro a variável de interesse. Então, uma ou mais variáveis adicionais são identificadas e controladas a fim de que se possam obter dados a respeito de como elas influem na variável de interesse. Por exemplo, uma empresa farmacêutica poderia estar interessada em realizar um experimento para saber como um novo medicamento afeta a pressão sangüínea. A pressão sangüínea é a variável de interesse no estudo. Espera-se que a dosagem do novo medicamento seja outra variável com efeito causal sobre a pressão sangüínea. Para obter dados sobre o efeito do novo medicamento, os pesquisadores selecionam uma amostra de indivíduos. A dosagem do novo medicamento é controlada, uma vez que diferentes grupos de pessoas recebem diferentes dosagens.

Antes e depois, coletam-se dados sobre a pressão sangüínea de cada um dos grupos. A análise estatística dos dados experimentais pode ajudar a determinar a maneira pela qual o novo medicamento afeta a pressão sangüínea.

Os estudos de fumantes e não-fumantes são baseados na informação porque os pesquisadores não determinam nem controlam quem fuma ou não.

Tabela 1.3 Exemplos de dados disponíveis em órgãos governamentais selecionados

Órgão Governamental	Dados Disponíveis
Census Bureau http://www.census.gov	Dados populacionais, número de famílias e renda familiar.
Federal Reserve Board http://www.federalreserve.gov	Dados sobre a base monetária, crédito de prestações, taxas de câmbio e taxas de desconto.
Office of Management and Budget http://www.whitehouse.gov/omb	Dados sobre a receita, gastos e débito do governo federal.
Department of Commerce http://www.doc.gov	Dados sobre a atividade empresarial, valor das exportações, nível de lucro da indústria e setores industriais que estão em crescimento ou declínio.
Bureau of Labor Statistics http://www.bls.gov	Gastos de consumo, remuneração por hora de trabalho, taxa de desemprego, registros de segurança no trabalho e estatísticas internacionais.

Figura 1.2 Página inicial do U.S. Census Bureau

FOIA | Privacy Statement | Confidentiality | Quality | Accessibility | Contact Us

USCENSUSBUREAU
Helping You Make Informed Decisions

Figura 1.3 Questionário de consulta aos clientes utilizado pelo Lobster Pot Restaurant, em Reddington Shores, Flórida

Estamos felizes com sua presença no Lobster Pot e queremos ter certeza de que você voltará a nos visitar. Assim, se tiver tempo, gostaríamos que preenchesse esta ficha. Seus comentários e sugestões são da máxima importância para nós. Muito obrigado!

Nome do garçom ______________________________

	Excelente	Boa	Satisfatória	Insatisfatória
Qualidade da comida	❑	❑	❑	❑
Gentileza no atendimento	❑	❑	❑	❑
Rapidez no atendimento	❑	❑	❑	❑
Higiene	❑	❑	❑	❑
Gerência	❑	❑	❑	❑

Comentários______________________________

____

O que o(a) motivou a nos visitar? ______________________________

____

Por favor, deposite esta ficha na caixa de sugestões na entrada do restaurante. Obrigado.

Estudos estatísticos não-experimentais, ou baseados na observação, não fazem nenhuma tentativa de controlar as variáveis de interesse. Uma pesquisa talvez seja o tipo mais comum de estudo baseado na observação. Por exemplo, em uma pesquisa que se realiza por meio de entrevistas pessoais, primeiramente são identificadas as perguntas a serem feitas. Depois, um questionário é projetado e ministrado a uma amostra de indivíduos. Alguns restaurantes utilizam estudos baseados na informação para obter dados sobre a opinião dos clientes quanto à qualidade da comida, atendimento, ambiente etc. Um questionário utilizado pelo Lobster Pot Restaurant, em Reddington Shores, Flórida, é apresentado na Figura 1.3. Observe que os clientes que respondem ao questionário são solicitados a apresentar avaliações de cinco variáveis: qualidade da comida, gentileza no atendimento, rapidez no atendimento, higiene e gerência. As categorias de resposta "excelente", "bom", "satisfatório" e "insatisfatório" fornecem dados ordinais que possibilitam aos gerentes do Lobster Pot avaliar a qualidade do funcionamento do restaurante.

Gerentes que queiram utilizar dados e análises estatísticas como apoio para a tomada de decisões devem estar cientes do tempo e custo necessários para a obtenção dos dados. O uso de fontes de dados existentes é desejável quando há a necessidade de os dados serem obtidos em um período relativamente curto. Se dados importantes não estiverem prontamente disponíveis, o tempo e o custo envolvidos em sua obtenção devem ser levados em conta. Em todos os casos, o tomador de decisões deve considerar a contribuição da análise estatística no processo de tomada de decisão. O custo da obtenção de dados e da subseqüente análise estatística não deve ultrapassar a economia gerada pelo uso da informação para se tomar uma decisão melhor.

Erros na Obtenção de Dados

Os gerentes devem sempre estar cientes da possibilidade de erros de dados nos estudos estatísticos. Usar dados errados pode ser pior do que não usar absolutamente nenhum dado. Um erro na obtenção de dados ocorre sempre que o valor de dados obtido não é igual ao valor verdadeiro ou real que seria obtido com um procedimento correto. Esses erros podem ocorrer de diversas maneiras. Por exemplo, um entrevistador poderia cometer um erro de registro, como a transposição ao escrever a idade de uma pessoa que tem 24 anos como uma de 42, ou a pessoa que responde às perguntas de uma entrevista poderia interpretar erroneamente a questão e fornecer uma resposta incorreta.

Analistas de dados experientes tomam muito cuidado ao coletar e registrar dados, a fim de assegurar que não se cometam erros. Procedimentos especiais podem ser usados para verificar a coerência interna dos dados. Por exemplo, esses procedimentos indicariam que o analista deve revisar a exatidão dos dados de uma pessoa que responde ter 22 anos de idade e 20 anos de experiência de trabalho. Os analistas de dados também revisam dados com valores incomumente elevados ou baixos, chamados dados "fora da curva", os quais são candidatos a possíveis erros. No Capítulo 3, apresentamos alguns dos métodos que os estatísticos usam para identificar esse tipo de dados.

Os erros freqüentemente ocorrem durante a obtenção dos dados. Utilizar cegamente quaisquer dados que possam estar disponíveis ou usar aqueles que foram obtidos com pouco cuidado pode resultar em informações enganosas e decisões ruins. Assim, tomar as medidas necessárias para obter dados precisos pode ajudar a assegurar que a informação será confiável e a tomada de decisões, valiosa.

1.4 ESTATÍSTICA DESCRITIVA

A maioria das informações estatísticas publicadas nos jornais, revistas, relatórios de empresas e outras publicações consiste em dados sintetizados e apresentados de forma fácil de entender para o leitor. Esses sumários de dados, que podem ser tabulares, gráficos ou numéricos, são conhecidos como **estatística descritiva**.

Consulte novamente o conjunto de dados da Tabela 1.1, que mostra dados referentes a 25 *shadow stocks*. Métodos de estatística descritiva podem ser usados para produzir sumários da informação contida nesse conjunto de dados. Por exemplo, um sumário tabular dos dados correspondentes à variável qualitativa bolsa de valores é exposto na Tabela 1.4. Um sumário gráfico dos mesmos dados encontra-se na Figura 1.4. Esses tipos de sumários tabulares e gráficos geralmente tornam os dados mais fáceis de ser interpretados. Consultando a Tabela 1.4 e a Figura 1.4, podemos ver facilmente que a maioria dos títulos do conjunto de dados é comercializada fora da bolsa (balcão). Em termos percentuais, 68% dos títulos são comercializados no balcão; 20%, na American Stock Exchange (Amex), e 20%, na New York Stock Exchange (Nyse).

Um sumário gráfico dos dados correspondentes à variável quantitativa Margem de Lucro Bruto dos *shadow stocks*, denominado histograma, é apresentado na Figura 1.5. No histograma torna-se fácil ver que as margens de lucro bruto variam de 0% a 75%, sendo as concentrações mais altas situadas entre 30% e 45%.

Tabela 1.4 Freqüências e freqüências percentuais da variável "bolsa de valores"

Bolsa de Valores	Freqüência	Freqüência Percentual
New York Stock Exchange (Nyse)	3	12
American Stock Exchange (Amex)	5	20
Over-the-counter (OTC)	17	68
Totais	25	100

Figura 1.4 Gráfico em barras da variável "bolsa de valores"

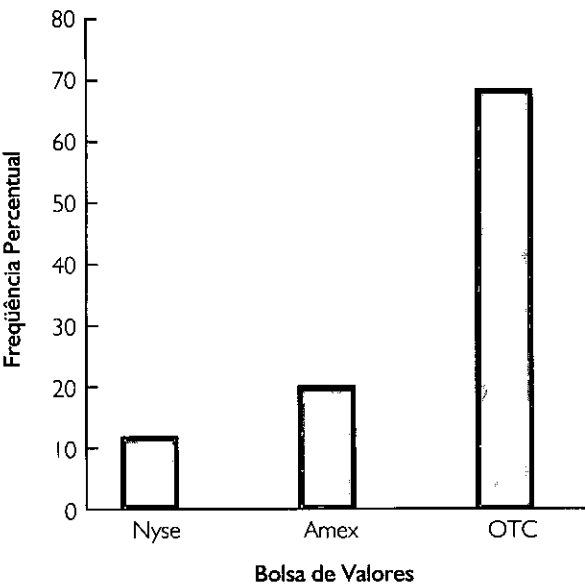
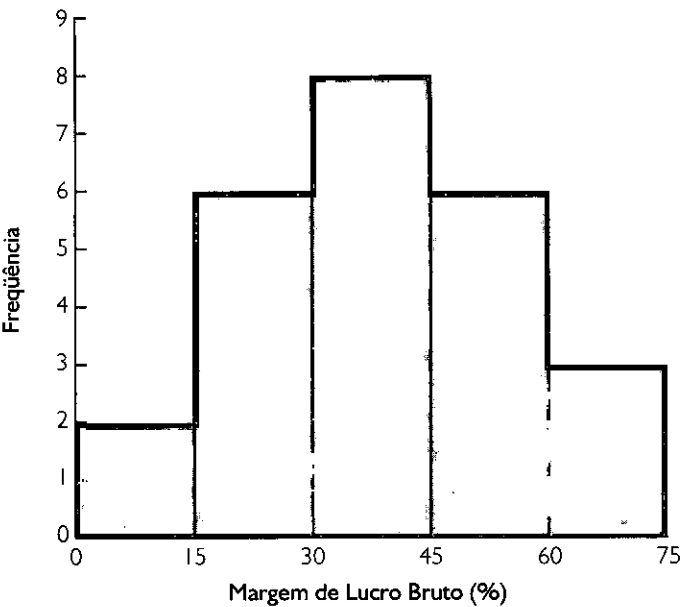



Figura 1.5 Histograma da margem de lucro bruto (%) dos 25 *shadow stocks*

Além das apresentações tabulares e gráficas, usam-se estatísticas descritivas numéricas para sintetizar os dados. A estatística descritiva numérica mais comum é a média. Usando os dados da variável capitalização de mercado dos *shadow stocks* mostrados na Tabela 1.1, podemos calcular a média de capitalização de mercado somando a capitalização de mercado de todas as 25 ações e dividindo a soma por 25. Essa operação produz uma média de capitalização de mercado igual a US$ 112,4 milhões. Essa média é tomada como uma medida da tendência central, ou posição central, dos dados correspondentes a essa variável.

Em muitas áreas, continua a aumentar o interesse nos métodos estatísticos que podem ser usados para desenvolver e apresentar estatísticas descritivas. Os Capítulos 2 e 3 dedicam atenção aos métodos estatísticos tabulares, gráficos e numéricos de estatística descritiva.

1.5 INFERÊNCIA ESTATÍSTICA

Muitas situações requerem dados relativos a um grupo amplo de elementos (indivíduos, empresas, eleitores, famílias, produtos, clientes etc.). Em virtude do tempo, custo e outros fatores, podem-se coletar dados somente de uma pequena parte do grupo. O grupo mais amplo dos elementos de determinado estudo denomina-se **população** e o grupo menor, **amostra**. Formalmente, usamos as seguintes definições:

O governo norte-americano realiza um censo a cada dez anos. As firmas de pesquisa de mercado realizam pesquisas amostrais todos os dias.

POPULAÇÃO
Uma população é o conjunto de todos os elementos de interesse em determinado estudo.

AMOSTRA
Uma amostra é um subconjunto da população.

O processo de realização de uma pesquisa para coletar dados correspondentes à população inteira chama-se **censo**. O processo de realização de uma pesquisa para coletar dados correspondentes a uma amostra denomina-se **pesquisa amostral**. Como uma de suas maiores contribuições, a estatística usa dados de uma amostra para fazer estimativas e testar hipóteses a respeito das características de uma população, utilizando um processo conhecido como **inferência estatística**.

Como um exemplo de inferência estatística, consideremos o estudo realizado pela Norris Electronics. A Norris produz um tipo de lâmpada de alta intensidade utilizada em uma série de produtos elétricos. Nesse caso, a população é definida como todas as lâmpadas que poderiam ser produzidas com o novo filamento. Para avaliar as vantagens do novo filamento, foram produzidas e testadas 200 lâmpadas com o novo filamento. Os dados coletados dessa amostra indicavam o número de horas que cada lâmpada permaneceu em operação antes de o filamento queimar-se. Veja a Tabela 1.5.

ARQUIVO DA INTERNET
Norris

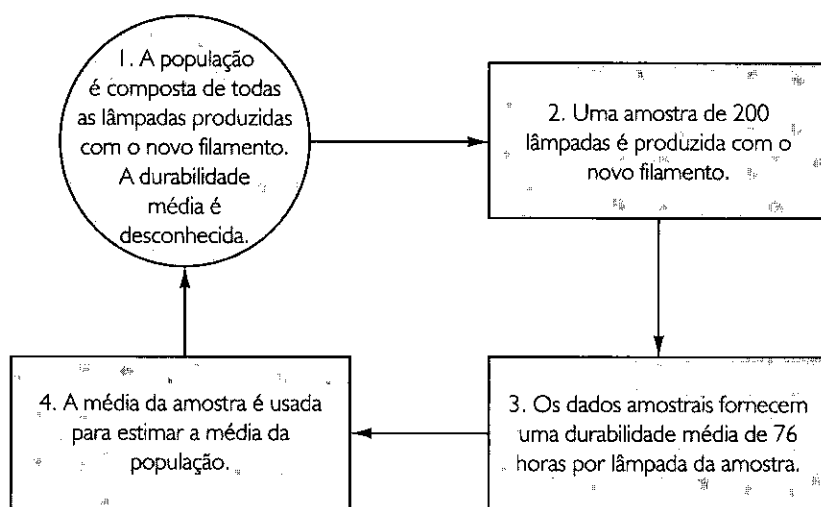
Tabela 1.5 A durabilidade em horas de uma amostra de 200 lâmpadas do exemplo da Norris Electronics

107	73	68	97	76	79	94	59	98	57
54	65	71	70	84	88	62	61	79	98
66	62	79	86	68	74	61	82	65	98
62	116	65	88	64	79	78	79	77	86
74	85	73	80	68	78	89	72	58	69
92	78	88	77	103	88	63	68	88	81
75	90	62	89	71	71	74	70	74	70
65	81	75	62	94	71	85	84	83	63
81	62	79	83	93	61	65	62	92	65
83	70	70	81	77	72	84	67	59	58
78	66	66	94	77	63	66	75	68	76
90	78	71	101	78	43	59	67	61	71
96	75	64	76	72	77	74	65	82	86
66	86	96	89	81	71	85	99	59	92
68	72	77	60	87	84	75	77	51	45
85	67	87	80	84	93	69	76	89	75
83	68	72	67	92	89	82	96	77	102
74	91	76	83	66	68	61	73	72	76
73	77	79	94	63	59	62	71	81	65
73	63	63	89	82	64	85	92	64	73

Suponha que a Norris queira usar os dados da amostra para fazer uma inferência a respeito da durabilidade média da população de todas as lâmpadas que poderiam ser produzidas com o novo filamento. A operação de somar os 200 valores da Tabela 1.5 e dividir o total por 200 produz a durabilidade média das lâmpadas da amostra: 76 horas. Podemos usar esse resultado amostral para estimar que a durabilidade média das lâmpadas da população é igual a 76 horas. A Figura 1.6 apresenta um sumário gráfico do processo de inferência estatística para a Norris Electronics.

Sempre que os estatísticos usam uma amostra para estimar determinada característica da população de interesse, geralmente apresentam uma declaração da qualidade, ou precisão, associada à estimativa.

Em relação ao exemplo da Norris, o estatístico poderia afirmar que a estimativa pontual da durabilidade média da população de novas lâmpadas é igual a 76 horas, com uma margem de erro de aproximadamente 4 horas. Assim, um intervalo estimado da durabilidade média para todas as lâmpadas produzidas é de 72 a 80 horas. O estatístico pode declarar também qual é o seu grau de confiança em que o intervalo de 72 a 80 horas contém a população média.

Figura 1.6 O processo de inferência estatística do exemplo da Norris Electronics

1.6 COMPUTADORES E A ANÁLISE ESTATÍSTICA

Uma vez que a análise estatística tipicamente envolve grandes quantidades de dados, os analistas freqüentemente usam software de computador para esse trabalho. Por exemplo, calcular a durabilidade média das 200 lâmpadas do exemplo da Norris Electronics (veja a Tabela 1.5) seria um trabalho bastante tedioso sem o uso de um computador. Para facilitar o uso do computador, os conjuntos de dados mais extensos deste livro estão disponíveis em www.thomsonlearning.com.br/estatapl.htm. Os arquivos de dados estão disponíveis tanto no formato Minitab como no formato Excel. Além disso, fornecemos instruções nos apêndices dos capítulos a respeito de como executar muitos dos procedimentos estatísticos usando o Minitab e o Excel.

Resumo

Estatística é a arte e ciência de coletar, analisar, apresentar e interpretar os dados. Quase todo estudante universitário que se especializa em negócios ou economia tem necessidade de fazer um curso de Estatística. Iniciamos o capítulo descrevendo as aplicações estatísticas típicas das áreas de administração e economia.

Dados são os fatos e os números que são coletados e analisados. As quatro escalas de medição usadas para obter dados sobre determinada variável são as seguintes: nominal, ordinal, intervalar e de proporção. A escala de medição de uma variável é nominal quando os dados utilizam rótulos ou nomes para identificar determinado atributo de um elemento. A escala é ordinal se os dados apresentam as propriedades inerentes aos dados nominais e a ordem, ou classificação, é significativa. A escala de medição é intervalar se os dados apresentam as propriedades inerentes aos dados ordinais e o intervalo entre os valores é expresso em termos

de uma unidade de medida fixa. Finalmente, a escala de medição é de proporção se os dados apresentam todas as propriedades inerentes aos dados de intervalo e a proporção dos dois valores é significativa.

Para fins de análise estatística, os dados podem ser classificados como qualitativos ou quantitativos. Os dados qualitativos usam rótulos ou nomes para identificar determinado atributo de cada elemento. Os dados qualitativos utilizam a escala de medição nominal ou a ordinal, e podem ser numéricos ou não-numéricos. Dados quantitativos são valores numéricos que indicam quantificação ou quantidade. Os dados quantitativos usam a escala de medição de intervalos ou de proporção. Operações numéricas comuns são significativas somente se os dados forem quantitativos. Portanto, cálculos estatísticos utilizados para dados quantitativos nem sempre são apropriados para dados qualitativos.

Nas Seções 1.4 e 1.5, apresentamos os tópicos da estatística descritiva e inferência estatística. Estatística descritiva são os métodos tabulares, gráficos e numéricos utilizados para sintetizar os dados. O processo de inferência estatística usa dados obtidos de uma amostra para fazer estimativas ou testar hipóteses referentes às características de uma população. Na última seção do capítulo destacamos que os computadores facilitam a análise estatística. Os conjuntos de dados mais extensos contidos em arquivos do Minitab ou Excel podem ser encontrados em www.thomsonlearning.com.br/estatapl.htm.

Glossário

Estatística A arte e ciência de coletar, analisar, apresentar e interpretar dados.

Dados Os fatos e os números que são coletados, analisados e sintetizados para apresentação e interpretação.

Conjunto de dados Todos os dados coletados em determinado estudo.

Elementos Entidades em relação às quais os dados são coletados.

Variável Característica dos elementos que nos interessa.

Observação Conjunto de medidas obtidas de dado elemento.

Escala nominal Escala de medição de uma variável quando os dados utilizam rótulos ou nomes para identificar determinado atributo de um elemento.

Escala ordinal Escala de medição de uma variável se os dados exibem as propriedades inerentes aos dados nominais e a ordem, ou classificação, dos dados é significativa. Os dados ordinais podem ser numéricos ou não-numéricos.

Escala intervalar Escala de medição de uma variável se os dados apresentam as propriedades inerentes aos dados ordinais e o intervalo entre os valores é expresso em termos de uma unidade de medida fixa. Os dados de intervalo são sempre numéricos.

Escala de proporção A escala de medição de uma variável se os dados demonstram todas as propriedades inerentes aos dados de intervalo e a proporção entre dois valores é significativa. Os dados de proporção são sempre numéricos.

Dados qualitativos Rótulos ou nomes usados para identificar um atributo de cada elemento. Os dados qualitativos utilizam a escala de medição nominal ou a ordinal e podem ser numéricos ou não-numéricos.

Dados quantitativos Valores numéricos que indicam a quantificação ou a quantidade de algo. Dados quantitativos são obtidos usando-se a escala de medição de intervalos ou de proporção.

Variável qualitativa Variável com dados qualitativos.

Variável quantitativa Variável com dados quantitativos.

Dados de seção transversal Dados coletados no mesmo ou aproximadamente no mesmo intervalo de tempo.

Dados de série histórica Dados coletados ao longo de diversos períodos.

Estatística descritiva Sumários tabulares, gráficos e numéricos de dados.

População Conjunto de todos os elementos que nos interessam em determinado estudo.

Amostra Subconjunto da população.

Censo Pesquisa com o objetivo de coletar dados sobre a população inteira.

Pesquisa amostral Uma pesquisa com o objetivo de coletar dados relativos a uma amostra.

Inferência estatística O processo de se usar os dados obtidos em uma amostra para fazer estimativas ou testar hipóteses a respeito das características de uma população.

Exercícios

1. Discuta a diferença entre a estatística como fatos numéricos e a estatística como disciplina ou área de estudo.

AUTOTESTE

2. A revista *Condé Nast Traveler* realiza uma pesquisa anual dos assinantes para determinar os melhores lugares para se hospedar em todos os lugares do mundo. A Tabela 1.6 apresenta uma amostra de nove hotéis europeus (*Condé Nast Traveler*, janeiro de 2000). O preço de um quarto de casal padrão durante a alta estação varia de $ (o preço mais baixo) a $$$$ (o preço mais alto). A pontuação global inclui a avaliação que os assinantes fazem dos quartos, do serviço, dos restaurantes, da localização/ambiente e das áreas públicas de cada hotel; uma pontuação global mais alta corresponde a um nível de satisfação mais elevado.

   a. Quantos elementos há nesse conjunto de dados?
   b. Quantas variáveis há nesse conjunto de dados?
   c. Quais variáveis são qualitativas e quais variáveis são quantitativas?
   d. Qual tipo de escala de medição é usada para cada uma das variáveis?

AUTOTESTE

3. Consulte a Tabela 1.6.

   a. Qual é o número médio de quartos dos nove hotéis?
   b. Calcule a pontuação global média.
   c. Qual é a porcentagem de hotéis localizados na Inglaterra?
   c. Qual é a porcentagem de hotéis com preços de quarto iguais a $$?

ARQUIVO DA INTERNET

Hotel

Tabela 1.6 Avaliações de nove lugares para se hospedar na Europa

Nome do Estabelecimento	País	Preço do Quarto	Número de Quartos	Pontuação Global
Graveteye Manor	Inglaterra	$$	18	83,6
Villa d'Este	Itália	$$$$	166	86,3
Hotel Prem	Alemanha	$	54	77,8
Hotel d'Europe	França	$$	47	76,8
Palace Luzern	Suíça	$$	326	80,9
Royal Crescent Hotel	Inglaterra	$$$	45	73,7
Hotel Sacher	Áustria	$$$	120	85,5
Duc de Bourgogne	Bélgica	$	10	76,9
Villa Gallici	França	$$	22	90,6

Fonte: Condé Nast Traveler, janeiro de 2000.

4. Os sistemas de som integrados, denominados *minisystems*, tipicamente incluem um rádio AM/FM, um *tape deck* duplo e um toca-CDs automático em um gabinete do tamanho de um livro com dois alto-falantes separados. Os dados da Tabela 1.7 mostram o preço de venda a varejo, a qualidade sonora, a capacidade de CDs, a sensibilidade e a seletividade do rádio AM/FM e o número de *tape decks* de uma amostra de 10 *minisystems* (*Consumer Report Buying Guide 2002*).

ARQUIVO DA INTERNET

Minisystems

Tabela 1.7 Uma amostra de dez *minisystems*

Marca e Modelo	Preço ($)	Qualidade Sonora	Capacidade de CDs	Sintonia de FM	*Tape Decks*
Aiwa NSX-AJ800	250	Boa	3	Razoável	2
JVC FS-SD1000	500	Boa	1	Muito Boa	0
JVC MX-G50	200	Muito Boa	3	Excelente	2
Panasonic SC-PM11	170	Razoável	5	Muito Boa	1
RCA RS 1283	170	Boa	3	Ruim	0
Sharp CD-BA2600	150	Boa	3	Boa	2
Sony CHC-CL1	300	Muito Boa	3	Muito Boa	1
Sony MHC-NX1	500	Boa	5	Excelente	2
Yamaha GX-505	400	Muito Boa	3	Excelente	1
Yamaha MCR-E100	500	Muito Boa	1	Excelente	0

a. Quantos elementos esse conjunto de dados contém?
b. Qual é a população?
c. Calcule o preço médio da amostra.
d. Usando os resultados obtidos no item (c), calcule a estimativa do preço médio da população.

5. Considere o conjunto de dados da amostra de dez *minisystems* da Tabela 1.7.
a. Quantas variáveis há no conjunto de dados?
b. Quais das variáveis são quantitativas e quais são qualitativas?
c. Qual é a capacidade média de CDs da amostra?
d. Qual porcentagem dos *minisystems* apresenta uma avaliação ótima ou excelente para sintonia de FM?
e. Qual porcentagem dos *minisystems* inclui dois *tape decks*?

6. A Columbia House entrega CDs aos membros do seu clube através de encomenda postal. A Columbia House Music Survey solicitou que os novos membros do clube preenchessem um formulário de pesquisa com 11 questões. Foram estas algumas das perguntas:
a. Quantos CDs você comprou nos últimos 12 meses?
b. Atualmente você é membro de algum "clube do livro" nacional que faz entregas por encomenda postal? (Sim ou Não).
c. Qual é a sua idade?
d. Quantas pessoas há em sua família (adultos e crianças), incluindo você?
e. Qual estilo de música você está interessado em comprar? (Foram relacionadas 15 categorias, incluindo *hard rock*, *soft rock*, música contemporânea adulta, *heavy metal*, *rap* e *country*.)
Comente se cada uma das perguntas fornece dados qualitativos ou quantitativos.

7. Uma pesquisa levada a efeito pela revista *Barron's* (15 de setembro de 2000) pediu aos assinantes para indicar qual era sua situação de emprego. Os dados foram registrados considerando uma escala em que 1 representava uma pessoa empregada em tempo integral, 2 representava uma pessoa empregada em tempo parcial, 3 representava uma pessoa aposentada e 4 representava alguém desempregado (dona de casa, estudante etc.).
a. A variável é a situação de emprego. Ela é uma variável qualitativa ou quantitativa?
b. Qual tipo de escala de medição é usado para essa variável?

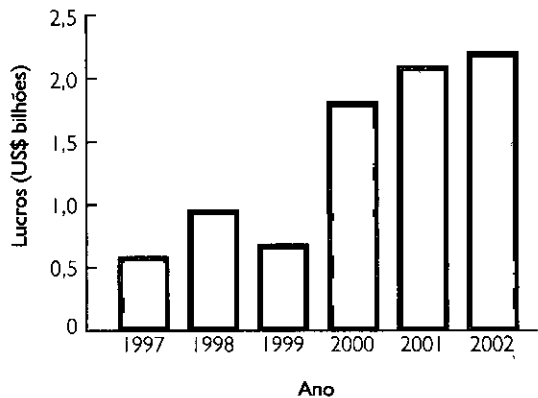
8. A organização Gallup realizou uma pesquisa telefônica com uma amostra nacional de 1.005 adultos escolhidos aleatoriamente, com idades a partir dos 18 anos. A pesquisa perguntou aos consultados: "Como você descreveria sua própria saúde neste momento?" (http://gallup.com, 7 de fevereiro de 2002). As categorias de resposta eram Excelente, Boa, Apenas Razoável, Ruim e Sem Opinião.
a. Qual foi o tamanho da amostra dessa pesquisa?
b. Os dados são qualitativos ou quantitativos?
c. Teria mais sentido usar médias ou porcentagens como um sumário dos dados para essa pergunta?
d. Dos consultados, 20% disseram que sua saúde pessoal estava excelente. Quantas pessoas deram essa resposta?

9. O Departamento do Comércio registrou as seguintes inscrições ao Prêmio Nacional da Qualidade Malcolm Baldrige (*Malcolm Baldrige National Quality Award*): 23 de grandes empresas de manufatura, 18 de grandes empresas de serviços e 30 de pequenos negócios.
a. O tipo de empresa é uma variável qualitativa ou quantitativa?
b. Qual porcentagem das aplicações veio de pequenos negócios?

10. Uma pesquisa do *The Wall Street Journal* (13 de outubro de 2003) entre seus assinantes contém 46 perguntas a respeito das características e interesses destes. Declare se cada uma das seguintes perguntas produziu dados qualitativos ou quantitativos e indique a escala de medição apropriada a cada uma:
a. Qual é a sua idade?
b. Você é do sexo masculino ou feminino?
c. Quando começou a ler o *Wall Street Journal*? No colégio, na universidade, no início da carreira, no meio da carreira, no fim da carreira, na aposentadoria?
d. Há quanto tempo você está em seu emprego ou cargo atual?
e. Qual tipo de veículo você pensa adquirir em sua próxima compra? Nove categorias de resposta incluem: sedã, carro esportivo, utilitário esportivo, minivan etc.

11. Declare se cada uma das seguintes variáveis é qualitativa ou quantitativa e indique sua escala de medição:
    a. Vendas anuais.
    b. Tamanho de refrigerante (pequeno, médio e grande).
    c. Classificação dos empregados (GS1 até GS18).
    d. Rendimento por ação.
    e. Método de pagamento (dinheiro, cheque, cartão de crédito).
12. O Hawaii Visitors Bureau coleta dados sobre visitantes que chegam ao Havaí. As perguntas apresentadas a seguir estão entre as 16 que foram formuladas em um questionário entregue aos passageiros das empresas aéreas que chegavam ao país, em junho de 2003.
    - Esta é a minha primeira, segunda, terceira, quarta etc. viagem ao Havaí.
    - O motivo principal para esta viagem é: (dez categorias, incluindo férias, convenções e lua-de-mel).
    - Onde planejo me hospedar: (11 categorias, incluindo hotel, apartamento, parentes, *camping*).
    - Tempo de permanência (em dias) no Havaí.

    a. Qual é a população estudada?
    b. A utilização de um questionário é uma boa maneira de atingir a população de passageiros que chegam nos vôos ao Havaí por via aérea?
    c. Comente cada uma das quatro perguntas em termos de elas fornecerem dados qualitativos e quantitativos.

Figura 1.7 Lucros da Volkswagen

13. A Figura 1.3 apresenta um gráfico de barras que sintetiza os lucros da Volkswagen correspondentes aos anos de 1997 a 2002 (*Business Week*, 23 de julho de 2001).
    a. Os dados são qualitativos ou quantitativos?
    b. Os dados são de série histórica ou transversais?
    c. Qual é a variável de interesse?
    d. Comente a tendência dos lucros da Volkswagem ao longo do tempo. Você esperaria ver uma elevação ou queda em 2003?

AUTOTESTE

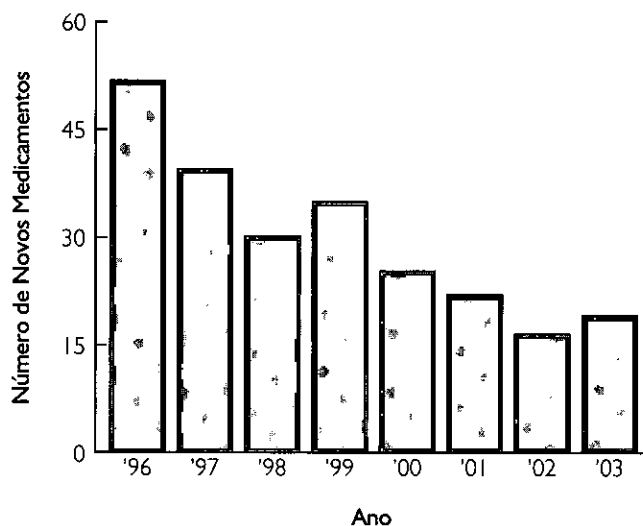
14. A Recording Industry of America faz um acompanhamento das vendas de gravações musicais levando em consideração o tipo de música, formato e faixa etária. Os dados a seguir apresentam as porcentagens das vendas de música de acordo com o tipo (*The New York Times 2002 Almanac*).

ARQUIVO DA INTERNET
Music

Tipo	1996	1997	1998	1999	2000
Rock	32,6	32,5	25,7	25,2	24,8
Country	12,1	11,2	12,8	10,8	10,7
R&B (*Rhythm and Blues*)	12,1	11,2	12,8	10,5	9,7
Pop	9,3	9,4	10,0	10,3	11,0
Rap	8,9	10,1	9,7	10,8	12,9
Gospel	4,3	4,5	6,3	5,1	4,8
Clássico	3,4	2,8	3,3	3,5	2,7
Jazz	3,3	2,8	1,9	3,0	2,9
Outros	14,0	15,5	17,5	20,8	20,5

a. O tipo de música é uma variável qualitativa ou quantitativa?
b. Construa um gráfico das vendas de *rock* ao longo de cinco anos; use o eixo horizontal para exibir o ano e o eixo vertical para exibir a porcentagem das vendas de gravações musicais. Este gráfico se baseia em dados de seção transversal ou de série histórica?
c. Construa um gráfico de barras do tipo de vendas musicais em 2000. Este gráfico se baseia em dados de seção transversal ou de série histórica?

15. A Food and Drug Administration (FDA) divulgou o número de novos medicamentos aprovados durante um período de oito anos (*The Wall Street Journal*, 12 de janeiro de 2004). A Figura 1.8 apresenta um gráfico de barras que sintetiza o número de novos medicamentos aprovados a cada ano.
a. Os dados são qualitativos ou quantitativos?
b. Os dados são de série temporal ou de seção transversal?
c. Quantos novos medicamentos foram aprovados em 2003?
d. Qual ano teve o menor número de medicamentos aprovados? Quantos?
e. Comente a tendência do número de novos medicamentos aprovados pela FDA no período de oito anos.

Figura 1.8 Número de novos medicamentos aprovados pela Food and Drug Administration (FDA)

16. A equipe de marketing de sua empresa desenvolveu um novo refrigerante dietético que, segundo afirmam, conquistará uma grande fatia do mercado para jovens e adultos.
a. Quais dados você quer examinar antes de decidir investir verbas substanciais para introduzir o novo produto no mercado?
b. Como você espera que os dados mencionados na questão (a) sejam obtidos?

17. Um gerente de uma grande corporação recomenda que seja dado um aumento salarial de US$ 10 mil a um subordinado valioso para impedi-lo de sair da empresa. Quais fontes internas e externas de dados poderiam ser usadas para decidir se esse aumento salarial é apropriado?

18. Uma pesquisa de 430 pessoas que viajam a negócios descobriu que 155 desses viajantes usavam um agente de viagens para fazer os arranjos da viagem (*USA Today*, 20 de novembro de 2003).
a. Desenvolva uma estatística descritiva que possa ser utilizada para estimar a porcentagem de todos os viajantes de negócios que usam um agente de viagens para fazer os arranjos da viagem.
b. A pesquisa divulgou que a maneira mais freqüente de os viajantes de negócios fazerem os arranjos de viagem é utilizando um site de viagens on-line. Se 44% dos viajantes de negócios pesquisados tiverem feito seus arranjos de viagem dessa maneira, quantos dos 430 viajantes de negócios usaram um site de viagens on-line?
c. Os dados relativos a como os arranjos de viagem são feitos são qualitativos ou quantitativos?

19. Um estudo dos assinantes norte-americanos da *Business Week* coletou dados de uma amostra de 2.861 assinantes. Cinqüenta e nove por cento dos entrevistados indicaram uma renda anual de US$ 75 mil ou mais, e 50% declararam ter um cartão de crédito American Express.
    a. Qual é a população de interesse nesse estudo?
    b. A renda anual é uma variável qualitativa ou quantitativa?
    c. Possuir um cartão de crédito American Express é uma variável qualitativa ou quantitativa?
    d. Esse estudo envolve dados de seção transversal ou de série histórica?
    e. Descreva quaisquer inferências estatísticas que a *Business Week* possa ter feito com base na pesquisa.

20. Um exame de 131 gerentes de investimentos que haviam participado da pesquisa de opinião Big Money da revista *Barron's* revelou o seguinte (*Barron's*, 28 de outubro de 2002):
    - 43% dos gerentes classificavam a si mesmos como especuladores otimistas (*bullish*) ou muito otimistas na bolsa de valores.
    - A média do valor esperado durantes os próximos 12 meses era de 11,2%.
    - 21% selecionaram o setor da saúde como o mais provável de liderar o mercado nos próximos 12 meses.
    - Quando solicitados a estimar quanto tempo seria necessário para que as ações de empresas de tecnologia e de comunicações retomassem um crescimento sustentável, a média da resposta dos gerentes foi 2,5 anos.

    a. Cite duas estatísticas descritivas.
    b. Faça uma inferência a respeito de toda a população de gerentes de investimento em relação à média de retorno esperado sobre o patrimônio líquido ao longo dos próximos 12 meses.
    c. Faça uma inferência a respeito da extensão de tempo necessária para que as ações das empresas de tecnologia e de telecomunicações retomem um crescimento sustentável.

21. O estudo de uma pesquisa médica de sete anos relatou que as mulheres cujas mães tomaram a droga DES (*dietilestilbestrol*) durante a gravidez tinham *o dobro* de probabilidade de desenvolver anormalidades celulares que poderiam resultar em câncer do que as mulheres cujas mães não haviam tomado.
    a. Esse estudo envolveu a comparação de duas populações. Quais eram elas?
    b. Você supõe que os dados foram obtidos em uma pesquisa ou em um experimento?
    c. Quanto à população de mulheres cujas mães tomaram a droga DES durante a gravidez, uma amostra de 3.980 mulheres apresentaram 63 anormalidades celulares que poderiam resultar em câncer. Forneça uma estatística descritiva que poderia ser usada para estimar o número de mulheres em cada grupo de mil dessa população que apresenta anormalidades celulares.
    d. Quanto à população de mulheres cujas mães não tomaram a droga DES durante a gravidez, qual é a estimativa do número de mulheres em cada grupo de mil que se poderia esperar que apresentem anormalidades celulares?
    e. Os estudos médicos freqüentemente usam uma amostra relativamente grande (nesse caso, 3.980). Por quê?

22. No verão de 2003, Arnold Schwarzenegger concorreu com o governador Gray Davis ao governo da Califórnia. Uma pesquisa realizada pelo Policy Institute of California entre os eleitores inscritos relatou que Arnold Schwarzenegger estava na liderança com uma estimativa de 54% dos votos (*Newsweek*, 8 de setembro de 2003).
    a. Qual foi a população dessa pesquisa?
    b. Qual foi a amostra dessa pesquisa?
    c. Por que foi usada uma amostra nessa situação? Explique.

23. A Nielsen Media Research realiza pesquisas semanais dos telespectadores em todo o território norte-americano e depois publica os dados de audiência e de fatia de mercado. A audiência relatada pela Nielsen é a porcentagem dos lares que possuem televisores e que estão assistindo a um programa, ao passo que a fatia de mercado é a porcentagem dos lares que assistem a um programa entre os lares que estão com o televisor ligado. Por exemplo, os resultados da Nielsen Media Research referentes ao Baseball World Series de 2003 entre o New York Yankees e o Florida Marlins mostraram uma audiência de 12,8% e uma fatia de mercado de 22% (Associated Press, 27 de outubro de 2003). Desse modo, 12,8% dos lares que possuem televisores estavam assistindo ao World Series e 22% dos lares que estavam com os televisores ligados assistiam ao World Series. Baseando-se nos dados de audiên-

cia e de fatia de mercado dos principais programas de televisão, a Nielsen publica uma classificação semanal desses programas, bem como uma classificação semanal das quatro principais redes: ABC, CBS, NBC e Fox.

a. O que a Nielsen Media Research tenta medir?
b. Qual é a população da pesquisa?
c. Por que uma amostra seria usada nessa situação?
d. Quais tipos de decisão ou ações se baseiam nas classificações da Nielsen?

24. Uma amostra de notas de cinco estudantes apresentou os seguintes resultados: 72, 65, 82, 90, 76. Quais das seguintes afirmações estão corretas e quais seriam contestadas como demasiadamente genéricas?

a. A nota média da amostra dos cinco estudantes é 77.
b. A nota média de todos os estudantes que fizeram o exame é 77.
c. Uma estimativa da nota média de todos os estudantes que fizeram o exame é 77.
d. Mais da metade dos estudantes que fizeram esse exame obterá pontos entre 70 e 85.
e. Se mais cinco estudantes forem incluídos na amostra, suas notas se situarão entre 65 e 90.

CAPÍTULO 2

Estatística Descritiva: Métodos Tabulares e Métodos Gráficos

ESTATÍSTICA NA PRÁTICA

A COMPANHIA COLGATE-PALMOLIVE*
Nova York, NY

A Companhia Colgate-Palmolive começou como uma pequena loja de sabões e velas em Nova York, em 1806. Hoje, a empresa emprega mais de 40 mil pessoas que trabalham em mais de 200 países e territórios mundo afora. Não obstante ser reconhecida internacionalmente por suas marcas Colgate, Palmolive e Ajax, a empresa também comercializa os produtos Mennen, Hill's Science Diet e Hill's Prescription Diet.

A Companhia Colgate-Palmolive utiliza a estatística em seu programa de garantia da qualidade para os produtos como detergentes de uso doméstico. Uma preocupação constante é a satisfação do cliente com a quantidade do produto na embalagem. Em cada categoria de tamanho, a embalagem é preenchida com a mesma quantidade de detergente em termos de peso, mas o volume do produto varia de acordo com a densidade do pó. Por exemplo, se a densidade do pó estiver mais concentrada, uma quantidade menor de detergente será necessária para atingir o peso especificado na caixa. Em conseqüência, a embalagem parecerá ter uma quantidade menor do produto quando for aberta pelo consumidor.

* Os autores agradecem a William R. Fowle, gerente de garantia da qualidade da companhia Colgate-Palmolive, por fornecer essa "Estatística na Prática".

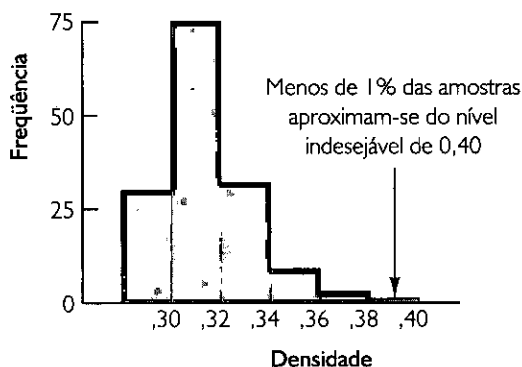
Para controlar o problema de peso do detergente, estabelecem-se os limites para o intervalo aceitável de densidade do pó. Amostras estatísticas são tomadas periodicamente, e a densidade de cada amostra de pó é medida. Dados sintetizados são então fornecidos ao pessoal do setor operacional a fim de que as ações corretivas possam ser tomadas, quando necessário, para manter a densidade dentro das especificações de qualidade desejadas.

A distribuição de freqüência e o histograma da densidade de 150 amostras tomadas no período de uma semana são exibidos na tabela e na figura que acompanham este texto. Níveis de densidade acima de 0,40 são inaceitavelmente altos. A distribuição de freqüência e o histograma revelam que a operação cumpre suas diretrizes de qualidade quando todas as densidades são menores ou iguais a 0,40. Gerentes que vissem esses sumários estatísticos ficariam satisfeitos com a qualidade do processo de produção do detergente.

Neste capítulo, você aprenderá os métodos tabulares e os métodos gráficos de estatística descritiva, como as distribuições de freqüência, gráficos em barras, histogramas, apresentações de ramo-e-folha, tabulações cruzadas e outros. O objetivo desses métodos é sintetizar os dados, de modo que eles possam ser facilmente entendidos e interpretados.

Distribuição de Freqüência dos Dados de Densidade

Densidade	Freqüência
0,29–0,30	30
0,31–0,32	75
0,33–0,34	32
0,35–0,36	9
0,37–0,38	3
0,39–0,40	1
Total	150

Histograma dos Dados de Densidade

Como vimos no Capítulo 1, os dados podem ser qualitativos ou quantitativos. Os **dados qualitativos** utilizam rótulos ou nomes para identificar categorias de itens semelhantes. Os **dados quantitativos** usam valores numéricos que indicam quantidade.

O propósito deste capítulo é apresentar os métodos tabulares e os métodos gráficos comumente usados para sintetizar tanto os dados qualitativos como os quantitativos. Sumários tabulares e gráficos de dados podem ser encontrados em relatórios anuais, artigos de jornais e em estudos de pesquisa. Esse tipo de apresentação nos é exposto com freqüência. Portanto, é importante entender como eles são elaborados e como devem ser interpretados. Abordaremos, em primeiro lugar, os métodos tabulares e os métodos gráficos para sintetizar dados referentes a uma única variável. A última seção introduz métodos para sintetizar dados quando a relação entre duas variáveis nos interessa.

Os modernos softwares estatísticos oferecem extensas capacidades para sintetizar dados e preparar apresentações gráficas. O Minitab e o Excel são dois pacotes de software amplamente disponíveis. Nos apêndices deste capítulo, mostraremos algumas de suas capacidades.

2.1 SINTETIZANDO OS DADOS QUALITATIVOS

A Distribuição de Freqüência

Iniciamos a discussão de como os métodos tabulares e os métodos gráficos podem ser usados para sintetizar os dados qualitativos com a definição de **distribuição de freqüência**.

DISTRIBUIÇÃO DE FREQÜÊNCIA
Uma distribuição de freqüência é um sumário tabular de dados que mostra o número (freqüência) de itens em cada uma das diversas classes não sobrepostas.

Vamos usar o exemplo seguinte para demonstrar a construção e interpretação de uma distribuição de freqüência correspondente aos dados qualitativos. Coca-Cola, Coca-Cola Light, Dr. Pepper, Pepsi-Cola e Sprite são cinco refrigerantes populares. Suponha que os dados da Tabela 2.1 mostrem o refrigerante selecionado em uma amostra de 50 compras de refrigerantes.

Tabela 2.1 Dados de uma amostra de 50 compras de refrigerantes

Coca-Cola	Sprite	Pepsi-Cola
Coca-Cola Light	Coca-Cola	Coca-Cola
Pepsi-Cola	Coca-Cola Light	Coca-Cola
Coca-Cola Light	Coca-Cola	Coca-Cola
Coca-Cola	Coca-Cola Light	Pepsi-Cola
Coca-Cola	Coca-Cola	Dr. Pepper
Dr. Pepper	Sprite	Coca-Cola
Coca-Cola Light	Pepsi-Cola	Coca-Cola Light
Pepsi-Cola	Coca-Cola	Pepsi-Cola
Pepsi-Cola	Coca-Cola	Pepsi-Cola
Coca-Cola	Coca-Cola	Pepsi-Cola
Dr. Pepper	Pepsi-Cola	Pepsi-Cola
Sprite	Coca-Cola	Coca-Cola
Coca-Cola	Sprite	Dr. Pepper
Coca-Cola Light	Dr. Pepper	Pepsi-Cola
Coca-Cola	Pepsi-Cola	Sprite
Coca-Cola	Coca-Cola Light	

ARQUIVO DA INTERNET
SoftDrink

Tabela 2.2 Distribuição de freqüência das compras de refrigerantes

Refrigerante	Freqüência
Coca-Cola	19
Coca-Cola Light	8
Dr. Pepper	5
Pepsi-Cola	13
Sprite	5
Total	50

Para desenvolver uma distribuição de freqüência desses dados, contamos o número de vezes que cada refrigerante aparece na Tabela 2.1. Coca-Cola aparece 19 vezes; Coca-Cola Light, oito; Dr. Pepper, cinco; Pepsi-Cola, 13 e Sprite, cinco vezes. Essas contagens encontram-se sintetizadas na distribuição de freqüência da Tabela 2.2.

Essa distribuição de freqüência fornece um resumo de como as 50 compras de refrigerantes estão distribuídas entre os cinco refrigerantes. Essa síntese fornece mais subsídios que os dados originais apresentados na Tabela 2.1. Observando a distribuição de freqüência, vemos que a Coca-Cola é a líder; a Pepsi-Cola, a segunda; a Coca-Cola Light, a terceira; e Sprite e Dr. Pepper estão empatados em quarto lugar. A distribuição de freqüência sintetiza informações sobre a popularidade dos cinco refrigerantes mais vendidos.

As Distribuições de Freqüência Relativa e de Freqüência Percentual

Uma distribuição de freqüência mostra o número (freqüência) de itens em cada uma das diversas classes não sobrepostas. Entretanto, muitas vezes estamos interessados na proporção, ou porcentagem, dos itens de cada classe. A *freqüência relativa* de uma classe equivale à fração ou proporção dos itens pertencentes a uma classe. Para um conjunto de dados com n observações, a freqüência relativa de cada classe pode ser determinada da seguinte maneira:

FREQÜÊNCIA RELATIVA

$$\text{Freqüência relativa de uma classe} = \frac{\text{Freqüência da classe}}{n} \qquad \textbf{(2.1)}$$

A *freqüência percentual* de uma classe é a freqüência relativa multiplicada por 100.

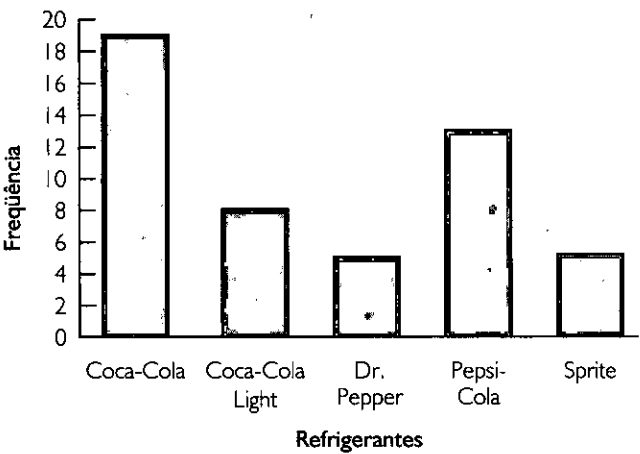
Uma **distribuição de freqüência relativa** constitui um sumário tabular de dados que mostra a freqüência relativa correspondente a cada classe. Uma **distribuição de freqüência percentual** sintetiza a freqüência percentual dos dados correspondentes a cada classe. A Tabela 2.3 mostra a distribuição de freqüência relativa e a distribuição de freqüência percentual dos refrigerantes. Nessa tabela, observamos que a freqüência relativa da Coca-Cola é 19/50 = 0,38, a freqüência relativa da Coca-Cola Light é 8/50 = 0,16 e assim por diante. A partir da distribuição de freqüência percentual, vemos que 38% das compras foram de Coca-Cola, 16% das compras foram de Coca-Cola Light etc. Podemos notar também que 38% + 26% + 16% = 80% das compras foram dos três principais refrigerantes.

Gráficos em Barras e em Setores ("Pizza")

Um **grafo de barras**, ou **gráfico em barras**, é um dispositivo gráfico para descrever os dados qualitativos que foram sintetizados em uma distribuição de freqüência, em uma distribuição de freqüência relativa ou em uma distribuição de freqüência percentual. Em um eixo do gráfico (geralmente, o eixo horizontal), especificamos os rótulos que são usados para as classes (categorias). Uma escala de freqüência, de freqüência relativa ou de freqüência percentual pode ser usada para o outro eixo do gráfico (normalmente, o eixo vertical).

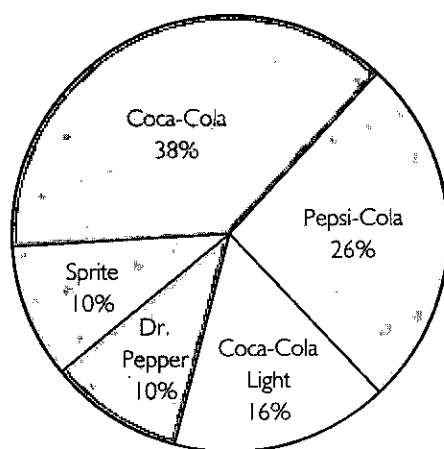
Tabela 2.3 Distribuições de freqüência relativa e de freqüência percentual das compras de refrigerantes

Refrigerante	Freqüência Relativa	Freqüência Percentual
Coca-Cola	0,38	38
Coca-Cola Light	0,16	16
Dr. Pepper	0,10	10
Pepsi-Cola	0,26	26
Sprite	0,10	10
Total	1,00	100

Figura 2.1 Gráfico em barras das compras de refrigerantes

Então, usando uma barra de largura fixa traçada acima de cada rótulo de classe, estendemos a altura da barra até atingirmos a freqüência, a freqüência relativa ou a freqüência percentual da classe. Para dados qualitativos, as barras devem estar separadas para enfatizar o fato de que cada classe é uma categoria distinta. A Figura 2.1 exibe um gráfico em barras da distribuição de freqüência correspondente às 50 compras de refrigerantes. Observe como a representação gráfica mostra que Coca-Cola, Pepsi-Cola e Coca-Cola Light são as marcas preferidas.

O **gráfico em setores ("pizza")** constitui outro dispositivo gráfico para representar as distribuições de freqüência relativa e as distribuições de freqüência percentual de dados qualitativos. Para construir um gráfico de pizza, traçamos primeiro um círculo para representar todos os dados. Depois, usamos as freqüências relativas para subdividir o círculo em setores, ou partes, que correspondem à freqüência relativa de cada classe. Por exemplo, uma vez que um círculo tem 360 graus e a Coca-Cola exibe uma freqüência relativa de 0,38, o setor do gráfico de pizza que detém o rótulo Coca-Cola consiste em $0{,}38 \times 360 = 136{,}8$ graus. O setor do gráfico de pizza que possui o rótulo Coca-Cola Light consiste em $0{,}16 \times 360 = 57{,}6$ graus. Cálculos idênticos para as outras classes produzem o gráfico em setores da Figura 2.2. Os valores numéricos mostrados para cada setor podem ser freqüências, freqüências relativas ou freqüências percentuais.

Em aplicações de controle da qualidade, os gráficos em barras são usados para identificar as causas mais importantes de problemas. Quando as barras são dispostas em ordem decrescente de altura, da esquerda para a direita, com a causa que tem a ocorrência mais freqüente aparecendo em primeiro lugar, o gráfico em barras denomina-se diagrama de Pareto. Esse diagrama é assim chamado em homenagem ao seu criador, Vilfredo Pareto, um economista italiano.

Figura 2.2 Gráfico em setores ("pizza") das compras de refrigerantes

NOTAS E COMENTÁRIOS

1. Freqüentemente o número de classes de uma distribuição de freqüência é igual ao número de categorias encontradas nos dados, como ocorre com os dados de compras de refrigerantes apresentados nesta seção. Os dados envolvem somente cinco marcas de refrigerantes, e uma classe de distribuição de freqüência distinta foi definida para cada uma delas. Dados que incluíssem todos os refrigerantes exigiriam muitas categorias, a maioria das quais teria um número muito pequeno de compras. Muitos estatísticos recomendam que as classes com freqüências menores sejam agrupadas em uma classe conjunta denominada "outros". Classes com freqüências iguais a 5% ou menos seriam, na maioria das vezes, tratadas dessa maneira.
2. A soma das freqüências em qualquer distribuição de freqüência sempre corresponde ao número de observações. A soma das freqüências relativas em qualquer distribuição de freqüência relativa sempre corresponde a 1,00, e a soma das porcentagens em uma distribuição de freqüência percentual sempre corresponde a 100.

Exercícios

Métodos

1. A resposta a uma questão tem três alternativas: A, B e C. Uma amostra de 120 respostas fornece 60 A, 24 B e 36 C. Mostre as distribuições de freqüência e de freqüência relativa.

2. Dada a seguinte distribuição de freqüência relativa:

Classe	Freqüência Relativa
A	0,22
B	0,18
C	0,40
D	

a. Qual é a freqüência relativa da classe D?
b. O tamanho total da amostra é 200. Qual é a freqüência da classe D?
c. Mostre a distribuição de freqüência.
d. Mostre a distribuição de freqüência percentual.

AUTOTESTE

3. Um questionário fornece as respostas: 58 sim, 42 não e 20 sem opinião.
a. Na construção de um gráfico de pizza, quantos graus teria a seção que representa as respostas afirmativas?
b. Quantos graus teria a seção do gráfico que apresenta as respostas negativas?
c. Construa um gráfico de pizza.
d. Construa um gráfico em barras.

Aplicações

ARQUIVO DA INTERNET
TVMedia

4. Os quatro programas de televisão de maior audiência nos Estados Unidos foram *CSI*, *ER*, *Everybody Loves Raymond* e *Friends* (*Nielsen Media Research*, 11 de janeiro de 2004). Seguem-se os dados que indicam os programas preferidos para uma amostra de 50 telespectadores:

CSI	Friends	CSI	CSI	CSI
CSI	CSI	Raymond	ER	ER
Friends	CSI	ER	Friends	CSI
ER	ER	Friends	CSI	Raymond
CSI	Friends	CSI	CSI	Friends
ER	ER	ER	Friends	Raymond
CSI	Friends	Friends	CSI	Raymond
Friends	Friends	Raymond	Friends	CSI
Raymond	Friends	ER	Friends	CSI
CSI	ER	CSI	Friends	ER

a. Esses dados são qualitativos ou quantitativos?
b. Forneça as distribuições de freqüência e de freqüência percentual.
c. Construa um gráfico em barras e um gráfico em setores ("pizza").
d. Com base na amostra, qual programa de televisão tem a maior audiência? Qual é o segundo colocado?

ARQUIVO DA INTERNET
Names

5. Em ordem alfabética, os seis sobrenomes mais comuns nos Estados Unidos são Brown, Davis, Johnson, Jones, Smith e Williams (*Time Almanac 2001*). Suponha que uma amostra de 50 indivíduos com um desses sobrenomes forneça os seguintes dados:

Brown	Williams	Williams	Williams	Brown
Smith	Jones	Smith	Johnson	Smith
Davis	Smith	Brown	Williams	Johnson
Johnson	Smith	Smith	Johnson	Brown
Williams	Davis	Johnson	Williams	Johnson
Williams	Johnson	Jones	Smith	Brown
Johnson	Smith	Smith	Brown	Jones
Jones	Jones	Smith	Smith	Davis
Davis	Jones	Williams	Davis	Smith
Jones	Johnson	Brown	Johnson	Davis

Sintetize os dados construindo o seguinte:
a. As distribuições de freqüência relativa e percentual.
b. Um gráfico em barras.
c. Um gráfico em setores.
d. Com base nesses dados, quais são os três sobrenomes mais comuns?

Tabela 2.4 Os oito livros de administração em capa flexível mais vendidos

- *The 7 Habits of Highly Effective People*
- *Investing for Dummies*
- *The Ernst & Young Tax Guide 2000*
- *The Millionaire Next Door*
- *The Motley Fool Investment Guide*
- *Rich Dad, Poor Dad*
- *The Wall Street Journal Guide to Understanding Money and Investing*
- *What Color is Your Parachute? 2000*

6. Os oito livros de Administração em capa mole mais vendidos estão relacionados na Tabela 2.4 (*Business Week*, 3 de abril de 2000). Suponha que uma amostra de compras de livros forneça os seguintes dados:

ARQUIVO DA INTERNET
BWBooks

7 Habits	Dad	7 Habits	Millionaire	Millionaire	WSJ Guide
Motley	Millionaire	Tax Guide	7 Habits	Dad	Dummies
Millionaire	Motley	Dad	Dad	Parachute	Dad
Dad	7 Habits	WSJ Guide	WSJ Guide	WSJ Guide	7 Habits
Motley	WSJ Guide	Millionaire	7 Habits	Millionaire	Millionaire
Millionaire	7 Habits	Millionaire	7 Habits	Motley	Motley
Motley	7 Habits	Dad	Dad	Dad	Dad
7 Habits	WSJ Guide	Tax Guide	Millionaire	Motley	Tax Guide
Motley	Motley	Millionaire	Millionaire	Dad	Dummies
Millionaire	Millionaire	Millionaire	Dad	Millionaire	Dad

a. Construa as distribuições de freqüência e de freqüência percentual desses dados. Agrupe os livros que têm uma freqüência igual a 5% ou menos em uma categoria denominada "outros".
b. Classifique os livros mais vendidos.
c. Quais porcentagens das vendas representam *The Millionaire Next Door* e *Rich Dad, Poor Dad*?

7. O Leverock's Waterfront Steakhouse, em Maderia Beach, Flórida, usa um questionário para perguntar aos clientes como eles avaliam o atendimento dos garçons, a qualidade das refeições, os drinques, os preços e o ambiente do restaurante. Cada característica é avaliada de acordo com uma escala que varia de excelente (E), ótimo (O), bom (B), médio (M) a fraco (F). Utilize a estatística descritiva para sintetizar os seguintes dados coletados sobre a qualidade das refeições. Qual é a sua impressão a respeito das avaliações da qualidade das refeições no restaurante?

AUTOTESTE

B	E	O	B	M	E	O	E	O	B	E	O	M
O	E	F	M	E	B	M	E	E	E	B	E	O
O	M	B	E	O	F	O	E	E	B	E	E	
E	B	M	E	O	E	E	B	O	M	B		

8. Os dados apresentados a seguir referem-se a uma amostra de 55 integrantes do Hall da Fama do Beisebol, em Cooperstown, Nova York. Cada observação indica a posição principal em que os integrantes do Hall da Fama jogavam: arremessador (A), receptor (R), primeira base (1), segunda base (2), terceira base (3), interbase (I), campo externo esquerdo (E), campo externo central (C) e campo externo direito (D).

E	A	C	R	2	A	D	1	I	I	1	E	A	D	A
A	A	A	D	C	I	E	D	A	C	C	A	A	D	A
2	3	A	R	E	A	1	C	A	A	A	I	1	E	D
D	1	2	R	I	3	R	2	E	A					

a. Use as distribuições de freqüência e de freqüência relativa para sintetizar os dados.
b. Qual posição contribui com mais integrantes para o Hall da Fama?
c. Qual posição contribui com menos integrantes para o Hall da Fama?
d. Qual posição de *outfield*[1] (E, C ou D) contribui com mais integrantes para o Hall da Fama?
e. Compare os *infielders*[2] (1, 2, 3 e I) com os *outfielders* (E, C e D).

[1] NT: *Outfield* – parte mais distante do campo, onde jogam os três jogadores (esquerda, centro e direita) (ou *outfielders*) (beisebol).
[2] NT: *Infielder* – jogadores que jogam no "diamante", ou seja, a parte mais próxima do campo delimitada pelas bases. Jogam nessa área o receptor, primeira, segunda e terceira bases, o interbase e o arremessador (beisebol).

9. Cerca de 60% dos negócios de pequeno e médio portes são negócios de família. Uma pesquisa realizada pela TEC International Inc. perguntou a *chief executive officers* (CEOs) de empresas familiares como eles se tornaram CEOs (*The Wall Street Journal*, 6 de dezembro de 2003). As respostas foram que o CEO herdou o negócio, o CEO construiu o negócio, ou o CEO foi contratado pela empresa familiar. Uma amostra de 26 CEOs de negócios de família forneceu os seguintes dados a respeito de como eles se tornaram CEOs.

ARQUIVO DA INTERNET

CEOs

Construiu	Construiu	Construiu	Herdou
Herdou	Construiu	Herdou	Construiu
Herdou	Construiu	Construiu	Construiu
Construiu	Foi contratado	Foi contratado	Foi contratado
Herdou	Herdou	Herdou	Construiu
Construiu	Construiu	Construiu	Foi contratado
Construiu	Herdou		

a. Forneça uma distribuição de freqüência.
b. Forneça uma distribuição de freqüência percentual.
c. Construa um gráfico em barras.
d. Qual porcentagem de CEOs de negócios de família tornaram-se CEOs porque herdaram a empresa? Qual é a principal razão para uma pessoa tornar-se o CEO de um negócio de família?

10. Uma pesquisa de satisfação do cliente realizada pela Merrill Lynch em 2001 solicitou aos clientes para indicarem quão satisfeitos eles estavam com seus serviços de consultoria financeira. As respostas dos clientes foram codificadas de 1 a 7, e 1 indicava "absolutamente em nada satisfeito" e 7, "extremamente satisfeito". Suponha que os dados a seguir sejam de uma amostra de 60 respostas referentes a um consultor financeiro em particular.

ARQUIVO DA INTERNET

Client

5	7	6	6	7	5	5	7	3	6
7	7	6	6	6	5	5	6	7	7
6	6	4	4	7	6	7	6	7	6
5	7	5	7	6	4	7	5	7	6
6	5	3	7	7	6	6	6	6	5
5	6	6	7	7	5	6	4	6	6

a. Explique por que esses dados são qualitativos.
b. Forneça uma distribuição de freqüência e uma distribuição de freqüência relativa desses dados.
c. Forneça um gráfico em barras.
d. Com base em seus sumários, comente a respeito da avaliação global que os clientes fazem do consultor financeiro.

2.2 SINTETIZANDO OS DADOS QUANTITATIVOS

A Distribuição de Freqüência

Conforme definimos na Seção 2.1, uma distribuição de freqüência é um sumário tabular de dados que mostra o número (freqüência) de itens em cada uma das diversas classes não sobrepostas. Essa definição vale tanto para os dados quantitativos como para os qualitativos. Entretanto, em relação aos dados quantitativos, devemos ser mais cuidadosos ao definir as classes não sobrepostas a serem usadas na distribuição de freqüência.

Por exemplo, considere os dados quantitativos apresentados na Tabela 2.5. Esses dados apresentam o tempo necessário, em dias, para serem concluídas as auditorias de fim de ano de uma amostra de 20 clientes da Sanderson and Clifford, uma pequena firma de contabilidade. As três etapas necessárias para definir as classes de uma distribuição de freqüência com dados quantitativos são:

1. Determinar o número de classes não sobrepostas.
2. Determinar a amplitude de cada classe.
3. Determinar os limites da classe.

Vamos demonstrar essas etapas desenvolvendo uma distribuição de freqüência dos dados de tempo para a conclusão das auditorias apresentados na Tabela 2.5.

ARQUIVO DA INTERNET
Audit

Tabela 2.5 Tempo (em dias) para a conclusão das auditorias de fim de ano

12	14	19	18
15	15	18	17
20	27	22	23
22	21	33	28
14	18	16	13

Construir as classes com a mesma amplitude reduz a possibilidade de interpretações inadequadas da parte do usuário.

Número de classes As classes são formadas especificando-se os intervalos que serão usados para agrupar os dados. Como diretriz geral, recomendamos usar entre 5 e 20 classes. Para um número pequeno de itens de dados, apenas cinco ou seis classes podem ser empregadas para sintetizar os dados. Para um número maior de itens de dados, geralmente é necessário um número maior de classes. A meta é usar classes suficientes para mostrar a variação nos dados, mas não tantas classes a ponto de algumas conterem somente alguns itens de dados. Uma vez que o número de itens de dados apresentados na Tabela 2.5 é relativamente pequeno ($n = 20$), optamos por desenvolver uma distribuição de freqüência com cinco classes.

Amplitude das classes A segunda etapa na construção de uma distribuição de freqüência para dados quantitativos é escolher uma amplitude para as classes. Como diretriz geral, recomendamos que a amplitude seja a mesma para cada uma das classes. Desse modo, a escolha do número de classes e a amplitude das classes não são decisões independentes. Um número maior de classes significa menor amplitude de classe e vice-versa. Para determinar uma amplitude de classe aproximada, começamos por identificar os maiores e os menores valores no conjunto de dados. Então, com o número desejado de classes especificado, podemos usar a seguinte expressão para estabelecer a amplitude aproximada da classe:

$$\text{Amplitude aproximada de classe} = \frac{\text{Maior valor entre os dados} - \text{Menor valor entre os dados}}{\text{Número de classes}} \quad \textbf{(2.2)}$$

A amplitude aproximada de classe fornecida pela Equação 2.2 pode ser arredondada para um valor mais conveniente, baseado na preferência da pessoa que desenvolve a distribuição de freqüência. Por exemplo, a amplitude aproximada de classe 9,28 poderia ser arredondada para 10, simplesmente porque 10 é uma classe mais conveniente de usar para representar uma distribuição de freqüência.

Não há distribuição de freqüência ideal para um conjunto de dados. Diferentes pessoas podem construir diferentes distribuições de freqüência igualmente aceitáveis. O objetivo é revelar o agrupamento natural e a variação dos dados.

Em relação ao conjunto de dados que envolve o tempo para a conclusão das auditorias de fim de ano, o maior valor é 33 e o menor, 12. Uma vez que decidimos sintetizar os dados com cinco classes, usando a Equação 2.2 obtemos uma amplitude aproximada de classe igual a $(33 - 12)/5 = 4{,}2$. Portanto, decidimos arredondar para cima e usar uma amplitude de classe de cinco dias na distribuição de freqüência.

Na prática, o número de classes e a amplitude aproximada de classe são determinados pelo método de tentativa-e-erro. Assim que um número possível de classes é escolhido, a Equação 2.2 é usada para se encontrar a amplitude aproximada de classe. O processo pode ser repetido para um número diferente de classes. Por fim, o analista utiliza o julgamento para determinar a combinação do número de classes e a amplitude de classe que provê a melhor distribuição de freqüência para sintetizar os dados.

Com relação aos dados de tempo para a conclusão das auditorias apresentados na Tabela 2.5, depois de decidirmos usar cinco classes, cada uma das quais com uma amplitude de cinco dias, a próxima tarefa é especificar os limites de classe para cada uma das cinco classes.

Limites de Classe Os limites de classe devem ser escolhidos de modo que cada uma das observações pertença a uma e somente uma classe. O *limite inferior de classe* identifica o menor valor de dados possível atribuído à classe. O *limite superior de classe* identifica o maior valor de dados possível atribuído à classe. Ao desenvolver distribuições de freqüências para dados qualitativos, não precisamos especificar os limites de classe porque cada item de dados situa-se naturalmente em uma classe distinta. Mas, quando se trata de dados quantitativos, por exemplo, os tempos para a conclusão das auditoria apresentados na Tabela 2.5, os limites de classe são necessários para determinar o lugar a que pertence cada valor de dados.

Usando os dados do tempo para a conclusão das auditorias apresentados na Tabela 2.5, escolhemos 10 dias como o limite inferior de classe e 14 dias como o limite superior de classe para a primeira classe. Essa

classe é denotada como 10-14 na Tabela 2.6. O menor valor de dados, 12, está incluído na classe 10-14. Escolhemos, então, 15 dias como o limite inferior de classe e 19 como o limite superior de classe para a classe seguinte. Continuamos a definir os limites inferiores e superiores para obtermos um total de cinco classes: 10-14, 15-19, 20-24, 25-29 e 30-34. O maior valor de dados, 33, está incluído na classe 30-34. A diferença entre os limites inferiores de classes adjacentes é a amplitude de classe. Usando os dois primeiros limites inferiores de classe, 10 e 15, vemos que a amplitude de classe é 15 – 10 = 5.

Uma vez determinados o número de classes, a amplitude de classe e os limites de classe, uma distribuição de freqüência pode ser obtida contando-se o número de valores de dados que pertencem a cada uma das classes. Por exemplo, os dados da Tabela 2.5 mostram que quatro valores – 12, 14, 14 e 13 – pertencem à classe 10-14. Desse modo, a freqüência para a classe 10-14 é 4. Prosseguindo com esse processo de contagem para as classes 15-19, 20-24, 25-29 e 30-34, obtemos a distribuição de freqüência da Tabela 2.6. Usando essa distribuição de freqüência, podemos observar o seguinte:

1. Os tempos mais freqüentes para a conclusão das auditorias encontram-se na classe de 15-19 dias. Oito dos 20 tempos para auditoria pertencem a essa classe.
2. Somente uma auditoria necessitou de mais de 30 dias.

Outras conclusões são possíveis, dependendo dos interesses da pessoa que visualiza a distribuição de freqüência. O mérito de uma distribuição de freqüência é que ela fornece *insights*[3] a respeito dos dados que não são facilmente obtidos quando se observam os dados em sua forma original não organizada.

Tabela 2.6 Distribuição de freqüência para os dados de tempo para a conclusão das auditorias

Tempo para a Conclusão das Auditorias (dias)	Freqüência
10–14	4
15–19	8
20–24	5
25–29	2
30–34	1
Total	20

Ponto médio da classe Em algumas aplicações, queremos conhecer os pontos médios das classes em uma distribuição de freqüência para os dados quantitativos. O **ponto médio da classe** é o valor intermediário entre os limites superior e inferior da classe. Para os dados de tempo para a conclusão das auditorias, os cinco pontos médios são 12, 17, 22, 27 e 32.

As Distribuições de Freqüência Relativa e de Freqüência Percentual

Definimos as distribuições de freqüência relativa e de freqüência percentual para os dados quantitativos da mesma maneira que o fazemos para os dados qualitativos. Primeiramente, lembre-se de que a freqüência relativa é a proporção das observações pertencentes a uma classe. Com n observações,

$$\text{Freqüência relativa de uma classe} = \frac{\text{Freqüência da classe}}{n}$$

a freqüência percentual de uma classe é a freqüência relativa multiplicada por 100.

Baseando-se nas freqüências de classe da Tabela 2.6, sendo $n = 20$, a Tabela 2.7 mostra a distribuição de freqüência relativa e a distribuição de freqüência percentual correspondente aos dados de tempo para a conclusão da auditoria. Observe que 0,40 das auditorias, ou seja, 40%, necessitaram de 15 a 19 dias. Somente 0,05 das auditorias, ou seja, 5%, necessitaram de 30 ou mais dias. Novamente, interpretações e *insights* adicionais podem ser obtidos usando-se a Tabela 2.7.

[3] NT: *Insight* – compreensão repentina, em geral intuitiva, de suas próprias atitudes e comportamentos, de um problema, de uma situação (psicologia).

Tabela 2.7 Distribuições de freqüência relativa e percentual dos dados de tempo para a conclusão das auditorias

Tempo para a Conclusão das Auditorias (dias)	Freqüência Relativa	Freqüência Percentual
10–14	0,20	20
15–19	0,40	40
20–24	0,25	25
25–29	0,10	10
30–34	0,05	5
Total	1,00	100

Gráficos de Dispersão Unidimensional

Um dos sumários de dados mais simples é o **gráfico de dispersão unidimensional** (***dot plot***). Um eixo horizontal mostra o intervalo dos dados. Cada valor é representado por um ponto localizado acima do eixo. A Figura 2.3 representa o gráfico de dispersão unidimensional dos dados de tempo para a conclusão das auditorias apresentados na Tabela 2.5. Os três pontos localizados acima de 18 no eixo horizontal indicam que um tempo de auditoria de 18 dias ocorreu três vezes. Os gráficos de dispersão unidimensional exibem os detalhes dos dados e são úteis para comparar a distribuição dos dados de duas ou mais variáveis.

Histograma

Uma apresentação gráfica bastante comum de dados quantitativos é o **histograma.** Esse sumário gráfico pode ser preparado para dados que foram anteriormente sintetizados em uma distribuição de freqüência, de freqüência relativa ou de freqüência percentual. Um histograma é construído colocando-se a variável de interesse no eixo horizontal, e a freqüência, a freqüência relativa ou a freqüência percentual no eixo vertical. A freqüência, a freqüência relativa ou a freqüência percentual de cada classe é mostrada desenhando-se um retângulo cuja base é determinada pelos limites da classe no eixo horizontal e cuja altura é a freqüência, a freqüência relativa ou a freqüência percentual correspondentes.

A Figura 2.4 corresponde a um histograma dos dados de tempo para a conclusão das auditorias. Observe que a classe que tem a maior freqüência é mostrada pelo retângulo que aparece acima da classe correspondente a 15-19 dias. A altura do retângulo revela que a freqüência dessa classe é 8. Um histograma da distribuição de freqüência relativa ou percentual desses dados se assemelharia ao histograma da Figura 2.4, excetuando-se que o eixo vertical seria rotulado com valores de freqüência relativa ou percentual.

Como mostra a Figura 2.4, os retângulos adjacentes de um histograma se tocam. Diferentemente de um gráfico em barras, um histograma não contém nenhuma separação natural entre os retângulos de classes adjacentes. Esse formato é a convenção habitual para os histogramas. Uma vez que as classes correspondentes aos dados de tempo para a conclusão das auditorias são estabelecidas como 10-14, 15-19, 20-24, 25-29 e 30-34, poderia parecer que há a necessidade de intervalos de uma unidade de 14 para 15, de 19 para 20, de 24 para 25 e de 29 para 30. Esses intervalos são eliminados quando se constrói um histograma. A eliminação dos intervalos entre as classes em um histograma dos dados de tempo para a conclusão das auditorias ajuda a mostrar que todos os valores entre o limite inferior da primeira classe e o limite superior da última classe são possíveis.

Figura 2.3 Gráfico de dispersão unidimensional (*dot plot*) dos dados de tempo para a conclusão das auditorias

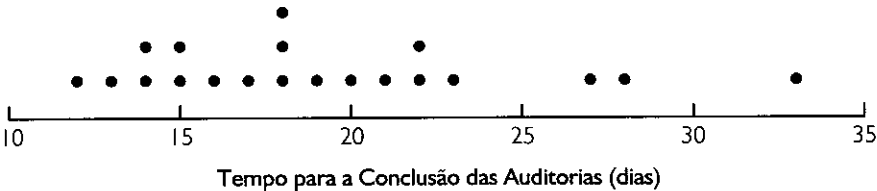
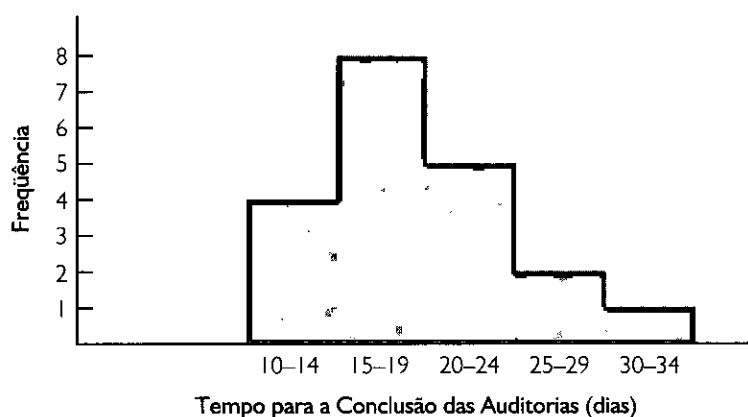



Figura 2.4 Histograma dos dados de tempo para a conclusão das auditorias

Uma das utilidades mais importantes de um histograma é fornecer informações sobre a forma, ou formato, de uma distribuição. A Figura 2.5 contém quatro histogramas construídos a partir de distribuições de freqüência relativa. O painel A mostra o histograma de um conjunto de dados moderadamente inclinado para a esquerda. Diz-se que um histograma é inclinado para a esquerda se sua cauda se estende bem à esquerda. Esse histograma é típico para a representação de pontuações obtidas em exames, com nenhuma pontuação acima de 100%, a maioria das pontuações acima de 70% e somente algumas pontuações realmente baixas. O painel B mostra o histograma de um conjunto de dados moderadamente inclinado para a direita. Diz-se que um histograma é inclinado para a direita se sua cauda se estende bem à direita. Um exemplo desse tipo de histograma seria o utilizado para representar dados como os preços de moradias; algumas casas muito caras criam a assimetria na cauda direita.

O painel C exibe um histograma simétrico. Em um histograma simétrico, a cauda esquerda espelha a forma da cauda direita. Histogramas para dados encontrados em aplicações jamais são perfeitamente simétricos, mas, para muitas aplicações, o histograma pode ser ligeiramente simétrico. Dados para pontuações no exame SAT,[4] altura e peso das pessoas etc. produzem histogramas ligeiramente simétricos. O painel D revela um histograma fortemente inclinado para a direita. Esse histograma foi construído a partir de dados sobre a quantidade de compras efetuadas por clientes no decorrer de um dia em uma loja de vestuário feminino. Dados de aplicações em negócios e economia freqüentemente produzem histogramas inclinados para a direita. Por exemplo, dados sobre preços de casas, salários, quantidade de compras etc. freqüentemente resultam em histogramas inclinados para a direita.

Distribuições Cumulativas

Uma variação da distribuição de freqüência que produz outro tipo de sumário tabular de dados quantitativos é a **distribuição de freqüência cumulativa**. A distribuição de freqüência cumulativa usa o número de classes, amplitudes de classes e limites de classe que foram desenvolvidas para a distribuição de freqüência. Entretanto, em vez de mostrar a freqüência de cada classe, a distribuição de freqüência cumulativa indica o número de itens de dados que possuem valores *menores ou iguais ao limite superior* de cada classe. As duas primeiras colunas da Tabela 2.8 apresentam a distribuição de freqüência cumulativa dos dados de tempo para a conclusão das auditorias.

Para entender como as freqüências cumulativas são determinadas, considere a classe com a descrição "menor ou igual a 24". A freqüência cumulativa dessa classe é simplesmente a soma das freqüências de todas as classes que possuem valores menores ou iguais a 24. Em relação à distribuição de freqüência da Tabela 2.6, a soma das freqüências correspondentes às classes 10-14, 15-19 e 20-24 indica que há 4 + 8 + 5 = 17 observações menores ou iguais a 24. Portanto, a freqüência cumulativa dessa classe é 17. Além disso, a distribuição de freqüência cumulativa apresentada na Tabela 2.8 indica que quatro auditorias foram concluídas em 14 dias ou menos, e que 19 auditorias foram concluídas em 29 dias ou menos.

---

[4] NT: O SAT (*Scholastic Aptitude Test*) é um exame usado pelas universidades norte-americanas como parte do processo de seleção de estudantes para a admissão ao curso superior; ele é realizado sete vezes por ano, envolvendo matemática e inglês.

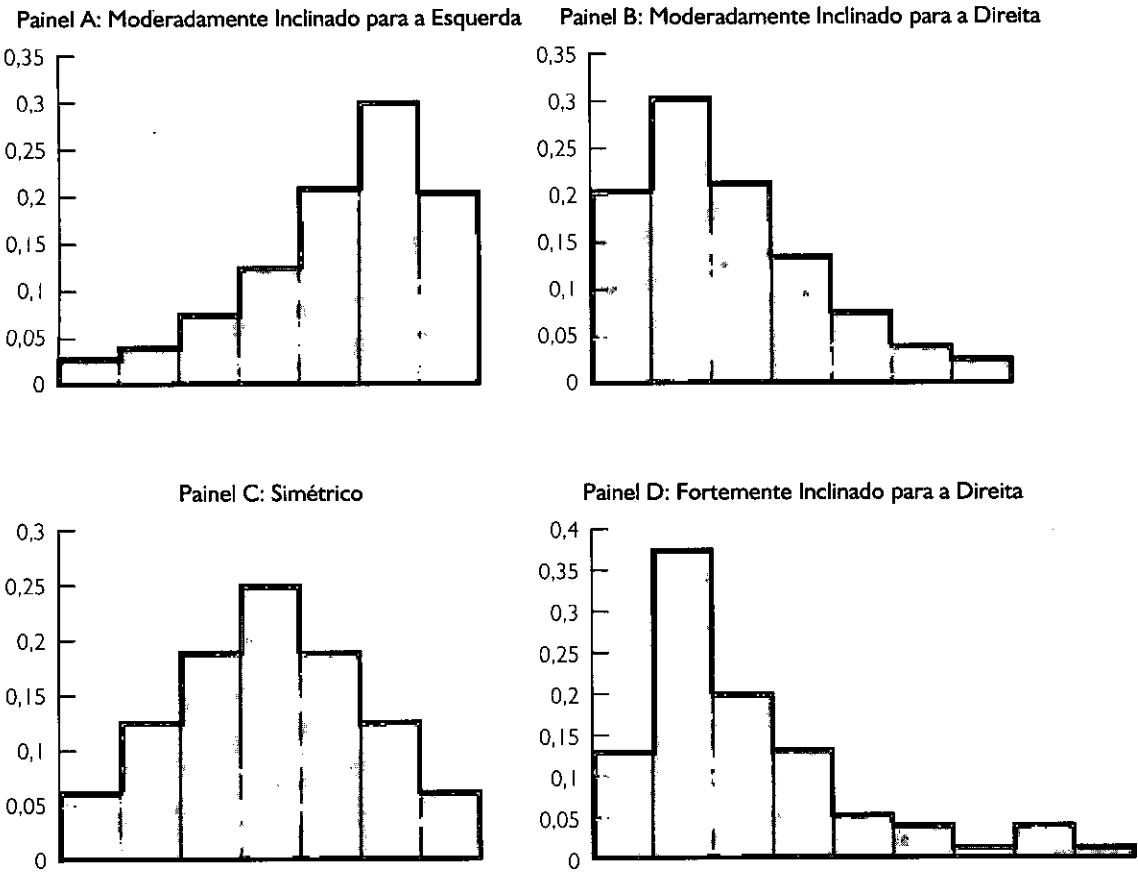
Figura 2.5 Histograma mostrando diferentes níveis de assimetria

Como observação final, notamos que uma **distribuição de freqüência relativa cumulativa** aponta a proporção de itens de dados, e que uma **distribuição de freqüência percentual cumulativa** mostra a porcentagem de itens de dados com valores menores ou iguais ao limite superior de cada classe. A distribuição de freqüência relativa cumulativa pode ser calculada somando-se as freqüências relativas existentes na distribuição de freqüência relativa ou dividindo-se as freqüências cumulativas pelo número total de itens. Utilizando a última abordagem, encontramos as freqüências relativas cumulativas na coluna 3 da Tabela 2.8 dividindo-se as freqüências cumulativas da coluna 2 pelo número total de itens (n = 20). As freqüências percentuais cumulativas foram novamente calculadas multiplicando-se as freqüências relativas por 100. As distribuições de freqüência relativa cumulativa e de freqüência percentual cumulativa mostram que 0,85 das auditorias, ou 85%, foram concluídas em 29 dias ou menos, 0,95 das auditorias, ou 95%, foram concluídas em 29 dias ou menos, e assim por diante.

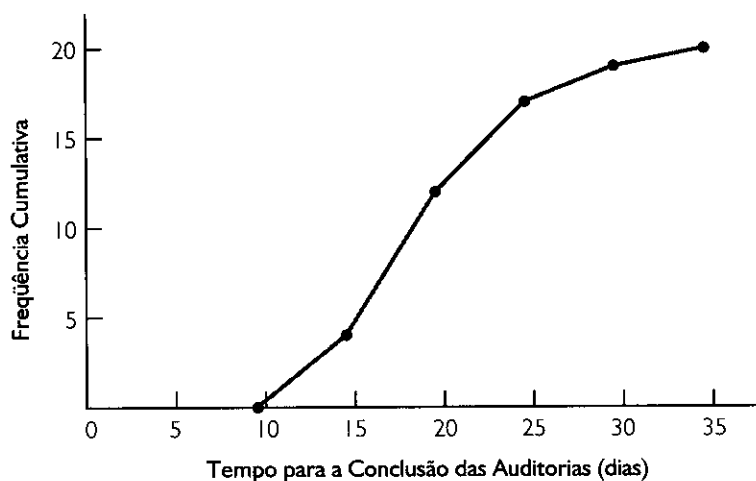
Tabela 2.8 Distribuições de freqüência cumulativa, de freqüência relativa cumulativa e de freqüência percentual cumulativa dos dados de tempo para a conclusão das auditorias

Tempo para a Conclusão das Auditorias (dias)	Freqüência Cumulativa	Freqüência Relativa Cumulativa	Freqüência Percentual Cumulativa
Menor ou igual a 14	4	0,20	20
Menor ou igual a 19	12	0,60	60
Menor ou igual a 24	17	0,85	85
Menor ou igual a 29	19	0,95	95
Menor ou igual a 34	20	1,00	100

Ogivas

O gráfico de uma distribuição cumulativa, chamado **ogiva**, mostra os valores dos dados no eixo horizontal e as freqüências cumulativas, as freqüências relativas cumulativas ou as freqüências percentuais cumulativas no eixo vertical. A Figura 2.6 ilustra uma ogiva correspondente às freqüências cumulativas dos dados de tempo para a conclusão das auditorias da Tabela 2.8.

A ogiva é construída assinalando-se um ponto correspondente à freqüência cumulativa de cada classe. Uma vez que as classes correspondentes aos dados de tempo para a conclusão das auditorias são 10-14, 15-19, 20-24 etc., intervalos de uma unidade aparecem de 14 para 15, de 19 para 20 e assim por diante. Esses intervalos são eliminados assinalando-se pontos intermediários entre os limites da classe. Desse modo, 14,5 é usado para a classe 10-14, 19,5 é usado para a classe 15-19 e assim por diante. A classe "menor ou igual a 14" com uma freqüência cumulativa igual a 4 é exposta na ogiva da Figura 2.6 pelo ponto localizado em 14,5 no eixo horizontal e 4 no eixo vertical. A classe "menor ou igual a 19" com uma freqüência cumulativa igual a 12 é indicada pelo ponto localizado em 19,5 no eixo horizontal e 12 no eixo vertical. Observe que um ponto adicional é assinalado na extremidade esquerda da ogiva. Esse ponto inicia a ogiva, mostrando que não há valores de dados abaixo da classe 10-14. Ele é assinalado em 9,5 no eixo horizontal e 0 no eixo vertical. Os pontos assinalados são conectados por linhas retas para preencher a ogiva.

Figura 2.6 Ogiva dos dados de tempo para a conclusão das auditorias

NOTAS E COMENTÁRIOS

1. Um gráfico em barras e um histograma são fundamentalmente iguais; ambos são representações gráficas dos dados em uma distribuição de freqüência. Um histograma é apenas um gráfico em barras sem nenhuma separação entre as barras. Para certos dados quantitativos discretos, uma separação entre as barras também é apropriada. Considere, por exemplo, o número de disciplinas nas quais um estudante universitário está matriculado. Os dados podem assumir somente valores inteiros. Valores intermediários, como 1,5; 2,73 etc., não são possíveis. Com dados quantitativos contínuos, entretanto, como os dados de tempo para a conclusão das auditorias da Tabela 2.5, uma separação entre as barras não é apropriada.
2. Os valores apropriados para os limites de classe com dados quantitativos dependem do nível de precisão dos dados. Por exemplo, com os dados de tempo para a conclusão das auditorias da Tabela 2.5 os limites usados foram valores inteiros. Se os dados fossem arredondados para o décimo de dia mais próximo (por exemplo, 12,3; 14,4 etc.), então os limites seriam declarados em décimos de dias. Por exemplo, a primeira classe seria 10,0-14,9. Se os dados fossem registrados para o centésimo de dia mais próximo (por exemplo, 12,34; 14,45 etc.), os limites seriam declarados em centésimos de dias. Por exemplo, a primeira classe seria 10,00-14,99.
3. Uma *classe aberta* requer somente um limite inferior de classe ou um limite superior de classe. Por exemplo, nos dados de tempo para a conclusão das auditorias apresentados na Tabela 2.5, suponha que

duas das auditorias tenham tomado 58 e 65 dias, respectivamente. Em vez de prosseguir com as classes de amplitude 5, como ocorre com as classes 35-39, 40-44, 45-49 etc., poderíamos simplificar a distribuição de freqüência para mostrar uma classe aberta de "35 ou mais". Essa classe teria uma freqüência igual a 2. Muito freqüentemente, a classe aberta aparece no lado superior da distribuição. Às vezes, uma classe aberta aparece no lado inferior da distribuição e, ocasionalmente, essas classes aparecem em ambos os lados.

4. A última entrada em uma distribuição de freqüência cumulativa sempre é igual ao número total de observações. A última entrada em uma distribuição de freqüência relativa cumulativa sempre é igual a 1,00 e a última entrada em uma distribuição de freqüência percentual cumulativa sempre é igual a 100.

Exercícios

Métodos

11. Considere os seguintes dados:

14	21	23	21	16
19	22	25	16	16
24	24	25	19	16
19	18	19	21	12
16	17	18	23	25
20	23	16	20	19
24	26	15	22	24
20	22	24	22	20

a. Desenvolva uma distribuição de freqüência usando classes de 12-14, 15-17, 18-20, 21-23 e 24-26.
b. Desenvolva uma distribuição de freqüência relativa e uma distribuição de freqüência percentual usando as classes apresentadas no item (a).

AUTOTESTE

12. Considere a seguinte distribuição de freqüência:

Classe	Freqüência
10-19	10
20-29	14
30-39	17
40-49	7
50-59	2

Construa uma distribuição de freqüência cumulativa e uma distribuição de freqüência relativa cumulativa.

13. Construa um histograma e uma ogiva dos dados do Exercício 12.

14. Considere os seguintes dados:

8,9	10,2	11,5	7,8	10,0	12,2	13,5	14,1	10,0	12,2
6,8	9,5	11,5	11,2	14,9	7,5	10,0	6,0	15,8	11,5

a. Construa um gráfico de dispersão unidimensional (*dot plot*).
b. Construa uma distribuição de freqüência.
c. Construa uma distribuição de freqüência percentual.

Aplicações

15. A equipe administrativa de um consultório médico estudou os tempos de espera dos pacientes que chegam ao consultório com um pedido de atendimento de emergência. Os seguintes dados de tempos de espera em minutos foram coletados no período de um mês:

2 5 10 12 4 4 5 17 11 8 9 8 12 21 6 8 7 13 18 3

Utilize classes de 0-4, 5-9 etc. para resolver as seguintes questões:

a. Mostre a distribuição de freqüência.
b. Mostre a distribuição de freqüência relativa.

c. Mostre a distribuição de freqüência cumulativa.
d. Mostre a distribuição de freqüência relativa cumulativa.
e. Qual proporção de pacientes que necessitam de atendimento de emergência enfrenta um tempo de espera de nove minutos ou menos?

16. Considere as seguintes distribuições de freqüência. A primeira distribuição de freqüência fornece uma aproximação da renda bruta ajustada anual nos Estados Unidos (Internal Revenue Service, março de 2003). A segunda distribuição de freqüência mostra as notas de exames dos estudantes de um curso universitário de Estatística.

Renda (US$ 1.000)	Freqüência (milhões)	Notas nos Exames	Freqüência
0–24	60	Abaixo de 30	2
25–49	33	30–39	5
50–74	20	40–49	6
75–99	6	50–59	13
100–124	4	60–69	32
125–149	2	70–79	78
150–174	1	80–89	43
175–199	1	90–99	21
Total	127	Total	200

a. Desenvolva um histograma dos dados de renda anual. Qual evidência de assimetria ele apresenta? Essa assimetria faz sentido? Explique.
b. Desenvolva um histograma dos dados de notas de exames. Qual evidência de assimetria ele apresenta? Explique.
c. Desenvolva um histograma dos dados do Exercício 11. Qual evidência de assimetria ele apresenta? Qual é a forma geral da distribuição?

ARQUIVO DA INTERNET
Spending

17. A Mendelsohn Media Research apresentou dados de pesquisa sobre a quantidade anual de compras domésticas feitas por famílias com uma renda anual de US$ 75.000 ou mais (*Money*, 2001). Suponha que os seguintes dados de uma amostra de 27 famílias indiquem a quantidade de dólares que elas gastaram no ano passado em livros e revistas.

280	496	382	202	287
266	119	10	385	135
475	255	379	267	24
42	25	283	110	423
160	123	16	243	363

a. Construa uma distribuição de freqüência e uma distribuição de freqüência relativa dos dados.
b. Forneça um histograma. Comente a respeito da forma da distribuição.
c. Comente a respeito dos gastos anuais em livros e revistas feitos pelas famílias da amostra.

ARQUIVO DA INTERNET
Wageweb

18. A Wageweb realiza pesquisas de dados salariais e apresenta os sumários em seu site. A empresa registrou que os salários anuais dos vice-presidentes de marketing variavam de US$ 85.090 a US$ 190.054 (Wageweb.com, 12 de abril de 2000). Suponha que os dados a seguir sejam de uma amostra dos salários anuais de 50 vice-presidentes de marketing. Os dados são em milhares de dólares:

145	95	148	112	132
140	162	118	170	144
145	127	148	165	138
173	113	104	141	142
116	178	123	141	138
127	143	134	136	137
155	93	102	154	142
134	165	123	124	124
138	160	157	138	131
114	135	151	138	157

a. Quais são os salários mais baixos e quais os mais altos?
b. Use uma amplitude de classe de US$ 15.000 e prepare sumários tabulares dos dados salariais anuais.
c. Qual proporção dos salários anuais são de US$ 135.000 ou menos?
d. Qual porcentagem dos salários anuais são superiores a US$ 150.000?
e. Prepare um histograma. Comente a respeito da forma da distribuição.

19. O trabalho de classificação de e-mails não-solicitados e *spam* afeta a produtividade de funcionários de escritório. Uma pesquisa levada a efeito pela InsightExpress monitorou funcionários de escritório para determinar a quantidade de tempo não-produtivo por dia dedicado a e-mails não-solicitados e *spam* (*USA Today*, 13 de novembro de 2003). Os dados a seguir fornecem uma amostra de tempo em minutos dedicado a essa tarefa:

2	4	8	4
8	1	2	32
12	1	5	7
5	5	3	4
24	19	4	14

Sintetize os dados construindo o seguinte:

a. Uma distribuição de freqüência (Classes 1-5, 6-10, 11-15, 16-20 etc.).
b. Uma distribuição de freqüência relativa.
c. Uma distribuição de freqüência cumulativa.
d. Uma distribuição de freqüência relativa cumulativa.
e. Uma ogiva.
f. Qual porcentagem de funcionários de escritório gasta cinco minutos ou menos em e-mails não-solicitados e *spam*? Qual porcentagem de funcionários de escritório gastam mais de dez minutos por dia nessa tarefa?

20. As 20 maiores turnês musicais e o preço médio dos ingressos de *shows* na América do Norte são mostrados a seguir. A lista baseia-se em dados fornecidos à publicação de negócios *Pollstar* por promotores de concertos e gerentes de eventos (*Associated Press*, 21 de novembro de 2003).

ARQUIVO DA INTERNET Concerts

Turnê Musical	Preço do Ingresso	Turnê Musical	Preço do Ingresso
Bruce Springsteen	$72,40	Toby Keith	$37,76
Dave Matthews Band	44,11	James Taylor	44,93
Aerosmith/Kiss	69,52	Alabama	40,83
Shania Twain	61,80	Harper/Johnson	33,70
Fleetwood Mac	78,34	50 Cent	38,89
Radiohead	39,50	Steely Dan	36,38
Cher	64,47	Red Hot Chili Peppers	56,82
Counting Crows	36,48	R.E.M.	46,16
Timberlake/Aguilera	74,43	American Idols Live	39,11
Mana	46,48	Mariah Carey	56,08

Sintetize os dados construindo o seguinte:

a. Uma distribuição de freqüência e uma distribuição de freqüência percentual.
b. Um histograma.
c. Qual concerto teve em média o preço de ingresso mais caro? Qual concerto teve em média o preço de ingresso mais barato?
d. Comente sobre o que os dados indicam a respeito da média dos preços de ingresso das maiores turnês musicais.

21. O *Nielsen Home Technology Report* apresentou informações sobre a tecnologia dos aparelhos domésticos e sua utilização por pessoas de 12 anos ou mais. Os dados a seguir referem-se ao número de horas de uso de computadores pessoais durante uma semana para uma amostra de 50 pessoas.

ARQUIVO DA INTERNET Computer

4,1	1,5	10,4	5,9	3,4	5,7	1,6	6,1	3,0	3,7
3,1	4,8	2,0	14,8	5,4	4,2	3,9	4,1	11,1	3,5
4,1	4,1	8,8	5,6	4,3	3,3	7,1	10,3	6,2	7,6
10,8	2,8	9,5	12,9	12,1	0,7	4,0	9,2	4,4	5,7
7,2	6,1	5,7	5,9	4,7	3,9	3,7	3,1	6,1	3,1

Sintetize os dados construindo o seguinte:

a. Uma distribuição de freqüência (use uma amplitude de classe de três horas).
b. Uma distribuição de freqüência relativa.
c. Um histograma.
d. Uma ogiva.
e. Comente sobre o que os dados indicam a respeito do uso de computadores pessoais em casa.

2.3 ANÁLISE EXPLORATÓRIA DOS DADOS: A APRESENTAÇÃO DE RAMO-E-FOLHA

As técnicas de **análise exploratória dos dados** consistem em cálculos aritméticos simples e em gráficos fáceis de desenhar que podem ser usados para sintetizar dados rapidamente. Uma dessas técnicas, denominada **apresentação de ramo-e-folha**, pode ser usada para mostrar simultaneamente tanto a ordem de classificação como a forma dos dados.

Para ilustrar o uso da apresentação de ramo-e-folha, considere os dados apresentados na Tabela 2.9. Esses dados resultam de um teste de aptidão composto de 150 questões aplicado a 50 pessoas entrevistadas recentemente para ocupar um cargo na Haskens Manufacturing. Os dados indicam o número de questões respondidas corretamente.

Para desenvolver uma apresentação de ramo-e-folha, organizamos primeiramente os dígitos à esquerda de cada valor de dados à esquerda de uma linha vertical. À direita da linha vertical, registramos o último dígito de cada valor de dados.

ARQUIVO DA INTERNET
Aptest

Tabela 2.9 Número de questões respondidas corretamente em um teste de aptidão

112	72	69	97	107
73	92	76	86	73
126	128	118	127	124
82	104	132	134	83
92	108	96	100	92
115	76	91	102	81
95	141	81	80	106
84	119	113	98	75
68	98	115	106	95
100	85	94	106	119

Com base na linha superior de dados da Tabela 2.9 (112, 72, 69, 97 e 107), as cinco primeiras entradas para se construir uma apresentação de ramo-e-folha seriam as seguintes:

6	9
7	2
8	
9	7
10	7
11	2
12	
13	
14	

Por exemplo, o valor de dados 112 mostra os dígitos à esquerda, 11, à esquerda da linha e o último dígito, 2, à direita da linha. De maneira similar, o valor de dados 72 mostra o dígito à esquerda, 7, à esquerda da linha e o último dígito, 2, à direita da linha. Continuando a colocar o último dígito de cada valor de dados na linha correspondente ao(s) seu(s) dígito(s) à esquerda, obtemos o seguinte:

```
 6 | 9 8
 7 | 2 3 6 3 6 5
 8 | 6 2 3 1 1 0 4 5
 9 | 7 2 2 6 2 1 5 8 8 5 4
10 | 7 4 8 0 2 6 6 0 6
11 | 2 8 5 9 3 5 9
12 | 6 8 7 4
13 | 2 4
14 | 1
```

Com essa organização dos dados, é simples classificar os dígitos de cada linha na devida ordem de classificação. Essa operação produz a apresentação de ramo-e-folha mostrada a seguir:

```
 6 | 8 9
 7 | 2 3 3 5 6 6
 8 | 0 1 1 2 3 4 5 6
 9 | 1 2 2 2 4 5 5 6 7 8 8
10 | 0 0 2 4 6 6 6 7 8
11 | 2 3 5 5 8 9 9
12 | 4 6 7 8
13 | 2 4
14 | 1
```

Os números à esquerda da linha vertical (6, 7, 8, 9, 10, 11, 12, 13 e 14) formam o *ramo*, e cada dígito à direita da linha vertical é uma *folha*. Por exemplo, considere a primeira linha com um valor de ramo 6 e folhas 8 e 9.

6 | 8 9

Isso indica que dois valores de dados têm um primeiro dígito 6. As folhas mostram que os valores de dados são 68 e 69. Similarmente, a segunda linha

7 | 2 3 3 5 6 6

indica que seis valores de dados têm um primeiro dígito 7. As folhas mostram que os valores de dados são 72, 73, 73, 75, 76 e 76.

Para nos concentrarmos na forma indicada na apresentação de ramo-e-folha, vamos usar um retângulo contendo as folhas de cada ramo. Com essa operação, obtemos o seguinte:

```
 6 | [8 9]
 7 | [2 3 3 5 6 6]
 8 | [0 1 1 2 3 4 5 6]
 9 | [1 2 2 2 4 5 5 6 7 8 8]
10 | [0 0 2 4 6 6 6 7 8]
11 | [2 3 5 5 8 9 9]
12 | [4 6 7 8]
13 | [2 4]
14 | [1]
```

Ao girarmos essa página 90 graus no sentido anti-horário, obtemos uma imagem dos dados que é similar a um histograma com as classes 60–69, 70–79, 80–89 e assim por diante.

Embora a apresentação de ramo-e-folha pareça oferecer a mesma informação dada por um histograma, ela tem duas vantagens principais:

1. A apresentação de ramo-e-folha é mais fácil de construir manualmente.
2. Dentro de um intervalo de classe, a apresentação de ramo-e-folha fornece mais informações que o histograma porque o ramo e a folha mostram os dados reais.

Da mesma forma que uma distribuição de freqüência ou um histograma não possuem um número absoluto de classes, também uma apresentação de ramo-e-folha não tem um número absoluto de linhas ou ramos. Se acharmos que nossa apresentação de ramo-e-folha condensou demasiadamente os dados, podemos facilmente estender a apresentação usando dois ou mais ramos para cada dígito à esquerda. Por exemplo, para usarmos dois ramos para cada dígito à esquerda, colocaríamos todos os valores de dados que terminam em 0, 1, 2, 3 e 4 em uma linha e todos os valores de dados que terminam em 5, 6, 7, 8 e 9 em uma segunda linha. A apresentação de ramo-e-folha estendida apresentada a seguir ilustra essa abordagem:

Em uma apresentação de ramo-e-folha estendida, sempre que um valor de ramo é declarado duas vezes, o primeiro valor corresponde aos valores de folha 0-4 e o segundo valor corresponde aos valores de folha 5-9.

6	8 9
7	2 3 3
7	5 6 6
8	0 1 1 2 3 4
8	5 6
9	1 2 2 2 4
9	5 5 6 7 8 8
10	0 0 2 4
10	6 6 6 7 8
11	2 3
11	5 5 8 9 9
12	4
12	6 7 8
13	2 4
13	
14	1

Note que os valores 72, 73 e 73 têm folhas no intervalo 0–4 e são mostrados com o primeiro valor de ramo 7. Os valores 75, 76 e 76 têm folhas no intervalo 5–9 e são mostrados no segundo valor de ramo 7. Essa apresentação de ramo-e-folha estendida é similar a uma distribuição de freqüência com intervalos 65–69, 70–74, 75–79 e assim por diante.

O exemplo anterior mostrou uma apresentação de ramo-e-folha de dados contendo até três dígitos. Apresentações de ramo-e-folha para dados com mais de três dígitos são possíveis. Por exemplo, considere os dados a seguir sobre o número de hambúrgueres vendidos por um restaurante de *fast-food* durante cada uma das 15 semanas:

1.565	1.852	1.644	1.766	1.888	1.912	2.044	1.812
1.790	1.679	2.008	1.852	1.967	1.954	1.733	

Uma apresentação de ramo-e-folha desses dados é a seguinte:

Unidade de folha = 10

15	6
16	4 7
17	3 6 9
18	1 5 5 8
19	1 5 6
20	0 4

Note que um único dígito é usado para definir cada folha e que somente os três primeiros dígitos de cada valor de dados foram usados para construir a apresentação. Na parte superior da apresentação, especifica-

mos a unidade de folha = 10. Para ilustrarmos a maneira de interpretar os valores da apresentação, considere o primeiro ramo, 15, e sua folha associada, 6. Combinando esses números, obtemos 156. Para reconstruirmos uma aproximação dos valores de dados originais, devemos multiplicar esse número por 10, que é o valor da *unidade de folha*. Desse modo, 156 × 10 = 1.560 é uma aproximação do valor de dados original utilizado para construir a apresentação de ramo-e-folha. Embora não seja possível reconstruir o valor de dados exato dessa apresentação de ramo-e-folha, a convenção de usar um único dígito para cada folha possibilita a construção de apresentações de ramo-e-folha para dados que contêm um número grande de dígitos. Para apresentações de ramo-e-folha em que a unidade de folha não é mostrada, presume-se que a unidade de folha seja igual a 1.

Um único dígito é usado para definir cada folha em uma apresentação de ramo-e-folha. A unidade de folha indica como multiplicar os números do ramo-e-folha a fim de obter uma aproximação dos dados originais. As unidades de folha podem ser 100, 10, 1, 0,1 e assim por diante.

Exercícios

Métodos

22. Construa uma apresentação de ramo-e-folha dos seguintes dados:

70	72	75	64	58	83	80	82
76	75	68	65	57	78	85	72

23. Construa uma apresentação de ramo-e-folha dos seguintes dados:

11,3	9,6	10,4	7,5	8,3	10,5	10,0
9,3	8,1	7,7	7,5	8,4	6,3	8,8

AUTOTESTE

24. Construa uma apresentação de ramo-e-folha dos seguintes dados. Use a unidade de folha 10.

1.161	1.206	1.478	1.300	1.604	1.725	1.361	1.422
1.221	1.378	1.623	1.426	1.557	1.730	1.706	1.689

Aplicações

25. Um psicólogo desenvolveu um novo teste de inteligência para adultos. O teste foi aplicado em 20 indivíduos, e os seguintes dados foram obtidos:

114	99	131	124	117	102	106	127	119	115
98	104	144	151	132	106	125	122	118	118

Construa uma apresentação de ramo-e-folha dos dados.

AUTOTESTE

26. A American Association of Individual Investors realiza uma pesquisa anual de *discount brokers*.[5] Os preços cobrados que são apresentados a seguir correspondem a uma amostra de 24 *discount brokers* (*AAII Journal*, janeiro de 2003). Os dois tipos de negócio são uma transação de 100 ações a US$ 50 por ação, o qual conta com a assistência de corretores, e um negócio on-line de 500 ações a US$ 50 por ação.

ARQUIVO DA INTERNET

Broker

Corretor	Negócio de 100 Ações a US$ 50 por Ação Auxiliado por Corretores	Negócio On-line de 500 Ações a US$ 50 por Ação	Corretor	Negócio de 100 Ações a US$ 50 por Ação Auxiliado por Corretores	Negócio On-line de 500 Ações a US$ 50 por Ação
Accutrade	30,00	29,95	Merrill Lynch Direct	50,00	29,95
Ameritrade	24,99	10,99	Muriel Siebert	45,00	14,95
Banc of America	54,00	24,95	NetVest	24,00	14,00
Brown & Co.	17,00	5,00	Recom Securities	35,00	12,95
Charles Schwab	55,00	29,95	Scottrade	17,00	7,00
CyberTrader	12,95	9,95	Sloan Securities	39,95	19,95
E*TRADE Securities	49,95	14,95	Strong Investments	55,00	24,95
First Discount	35,00	19,75	TD Waterhouse	45,00	17,95
Freedom Investments	25,00	15,00	T. Rowe Price	50,00	19,95
Harrisdirect	40,00	20,00	Vanguard	48,00	20,00
Investors National	39,00	62,50	Wall Street Discount	29,95	19,95
MB Trading	9,95	10,55	York Securities	40,00	36,00

[5] NT: *Discount broker* – As corretoras chamadas *discount broker*, ou de descontos, oferecem serviço de operação (compra e venda de futuros e opções da bolsa de valores) com foco na agilidade e na prática de preços. Elas apenas executam as ordens dos clientes, sem análise de papéis (economia).

a. Arredonde os preços de compra e venda para o valor em dólares mais próximo e desenvolva uma apresentação de ramo-e-folha das 100 ações a US$ 50 por ação. Comente sobre o que aprendeu a respeito dos preços da transação auxiliada por corretores.
b. Arredonde os preços da transação para o valor em dólares mais próximo e desenvolva uma apresentação de ramo-e-folha estendida das 500 ações on-line a US$ 50 por ação. Comente sobre o que aprendeu a respeito dos preços do negócio on-line.

27. Os preços por ação das 30 empresas que compõem a Dow Jones Industrial Average (*Média Industrial Dow Jones*) são mostrados a seguir (*The Wall Street Journal*, 9 de abril de 2004):

ARQUIVO
DA INTERNET
StockPrices

Empresa	Preço (em dólares) por ação	Empresa	Preço (em dólares) por ação
Alcoa	$34	Honeywell	$35
Altria Group	55	IBM	93
American Express	52	Intel	27
American International	76	Johnson & Johnson	51
Boeing	41	J.P. Morgan Chase	41
Caterpillar	82	McDonald's	29
Citigroup	52	Merck	45
Coca-Cola	51	Microsoft	25
Disney	26	Pfizer	36
DuPont	43	Procter & Gamble	106
ExxonMobil	42	SBE Communications	24
General Electric	31	3M	82
General Motors	47	United Technologies	90
Hewlett-Packard	23	Verizon	37
Home Depot	36	Wal-Mart	57

a. Desenvolva uma apresentação de ramo-e-folha.
b. Use a apresentação de ramo-e-folha para responder às seguintes questões:
   - O que o agrupamento dos dados da apresentação de ramo-e-folha lhe diz a respeito dos preços por ação das 30 empresas que compõem a Dow Jones?
   - Qual é a faixa de preço por ação da maioria das empresas?
   - Quantas empresas têm o preço de US$ 36 por ação?
   - Qual é o preço por ação que aparece mais freqüentemente?
   - Qual preço por ação seria considerado relativamente elevado? Qual porcentagem de empresas têm preços por ação nessa faixa? Quais empresas têm preços por ação nessa faixa e qual é o preço por ação de cada uma?
c. Use *The Wall Street Journal* ou outra publicação de negócios para descobrir o preço atual por ação de cada uma das 30 empresas que compõem a Dow Jones Industrial Average. Construa uma apresentação de ramo-e-folha desses dados e use a apresentação para comentar a respeito de quaisquer alterações nos preços por ação desde abril de 2004.

28. A minimaratona (20,92 km) de 2004, em Naples, Flórida, contou com 1.228 inscritos (*The Naples Daily News*, 17 de janeiro de 2004). A competição foi realizada em seis grupos distribuídos por faixa etária. Os dados a seguir mostram as idades de uma amostra de 40 indivíduos que participaram da maratona.

ARQUIVO
DA INTERNET
Marathon

49	33	40	37	56
44	46	57	55	32
50	52	43	64	40
46	24	30	37	43
31	43	50	36	61
27	44	35	31	43
52	43	66	31	50
72	26	59	21	47

a. Mostre uma apresentação de ramo-e-folha estendida.
b. Qual grupo etário teve o maior número de corredores?
c. Qual idade ocorreu mais freqüentemente?
d. Um artigo publicado no *Naples Daily News* destacou o número de corredores que "estavam na faixa etária dos 20 anos". Qual porcentagem dos corredores estava "na faixa etária dos 20 anos"? Na sua opinião, qual era o foco do artigo?

2.4 TABULAÇÕES CRUZADAS E DIAGRAMAS DE DISPERSÃO

Até agora, neste capítulo, focalizamos os métodos tabulares e gráficos utilizados para sintetizar os dados de *uma variável a cada vez*. Freqüentemente os gerentes ou tomadores de decisões necessitam de métodos tabulares e gráficos que lhes ajudem a compreender a *relação entre duas variáveis*. As tabulações cruzadas e os diagramas de dispersão são dois desses métodos.

Tabulação Cruzada

A **tabulação cruzada** é um sumário tabular de dados para duas variáveis. Vamos ilustrar a utilização de uma tabulação cruzada considerando a seguinte aplicação baseada em dados da Zagat's Restaurant Review. Dados sobre a avaliação da qualidade do restaurante e o preço das refeições foram coletados de uma amostra de 300 restaurantes localizados na região de Los Angeles. A Tabela 2.10 mostra os dados referentes aos dez primeiros restaurantes. Dados sobre a avaliação da qualidade do restaurante e preço típico das refeições são apresentados. A avaliação da qualidade é uma variável qualitativa com categorias de avaliação bom, ótimo e excelente. O preço das refeições é uma variável quantitativa que geralmente vai de US$ 10 a US$ 49.

As tabulações cruzadas e os diagramas de dispersão são usados para sintetizar dados de maneira que revele a relação entre duas variáveis.

Uma tabulação cruzada dos dados dessa aplicação é exibida na Tabela 2.11. Os rótulos das margens esquerda e superior definem as classes das duas variáveis. Na margem esquerda, os rótulos das linhas (bom, ótimo e excelente) correspondem às três classes da variável avaliação da qualidade. Na margem superior, os rótulos das colunas (US$ 10-19, US$ 20-29, US$ 30-39 e US$ 40-49) correspondem às quatro classes da variável preço das refeições. Cada restaurante da amostra apresenta uma avaliação da qualidade e o preço de uma refeição.

Desse modo, cada restaurante da amostra está associado a uma célula que aparece em uma das linhas e em uma das colunas da tabulação cruzada. Por exemplo, o restaurante 5 é identificado como aquele que tem uma avaliação de qualidade ótima e preço das refeições igual a US$ 33. Esse restaurante pertence à célula da linha 2, coluna 3, da Tabela 2.11. Ao construir uma tabulação cruzada, simplesmente contamos o número de restaurantes que pertencem a cada uma das células existentes na tabela de tabulação cruzada.

Tabela 2.10 Avaliação da qualidade e preço das refeições de 300 restaurantes de Los Angeles

Restaurante	Avaliação da Qualidade	Preço das Refeições (US$)
1	Bom	18
2	Ótimo	22
3	Bom	28
4	Excelente	38
5	Ótimo	33
6	Bom	28
7	Ótimo	19
8	Ótimo	11
9	Ótimo	23
10	Bom	13
•	•	•
•	•	•
•	•	•

ARQUIVO DA INTERNET
Restaurant

Revendo a Tabela 2.11, observamos que a maioria dos restaurantes da amostra (64) tem uma avaliação ótima e preço de refeições na faixa de US$ 20-29. Somente dois restaurantes têm uma avaliação excelente e preço de refeições na faixa de US$ 10-19. Podem ser feitas interpretações idênticas das outras freqüências. Além disso, observe que as margens direita e inferior da tabulação cruzada apresentam separadamente as distribuições de freqüência relativas à avaliação da qualidade do restaurante e preço das refeições. Da distribuição de freqüência na margem direita, notamos que os dados sobre as avaliações da qualidade mostram 84 restaurantes bons, 150 restaurantes ótimos e 66 restaurantes excelentes. Similarmente, a margem inferior mostra a distribuição de freqüência da variável preço das refeições.

Dividindo os totais indicados na margem direita da tabulação cruzada pelo total correspondente a essa coluna, obtemos uma distribuição de freqüência relativa e percentual da variável avaliação da qualidade.

Avaliação da Qualidade	Freqüência Relativa	Freqüência Percentual
Bom	,28	28
Ótimo	,50	50
Excelente	,22	22
Total	1,00	100

Tabela 2.11 Tabulação cruzada da avaliação da qualidade e preço das refeições de 300 restaurantes de Los Angeles

	Preço das Refeições				
Avaliação da Qualidade	$10–19	$20–29	$30–39	$40–49	Total
Bom	42	40	2	0	84
Ótimo	34	64	46	6	150
Excelente	2	14	28	22	66
Total	78	118	76	28	300

Da distribuição de freqüência percentual, observamos que 28% dos restaurantes foram avaliados como bons, 50% foram avaliados como ótimos e 22% foram avaliados como excelentes.

Ao dividir os totais da linha inferior da tabulação cruzada pelo total correspondente a essa linha, obtemos as distribuições de freqüência relativa e percentual da variável preço das refeições.

Preço da Refeição	Freqüência Relativa	Freqüência Percentual
$10–19	0,26	26
$20–29	0,39	39
$30–39	0,25	25
$40–49	0,09	9
Total	1,00	100

Note que a soma dos valores de cada coluna não coincide exatamente com o total da coluna, porque os valores somados são arredondados. Da distribuição de freqüência percentual, notamos que 26% dos preços das refeições encontram-se na classe de menor preço (US$ 10-19), 39% encontram-se na classe de preço mais elevado e assim por diante.

As distribuições de freqüência e de freqüência relativa construídas a partir das margens de uma tabulação cruzada fornecem informações a respeito de cada uma das variáveis individualmente, mas nada esclarecem a respeito da relação entre as variáveis. O principal mérito de uma tabulação cruzada reside no *insight* que ela oferece a respeito da relação entre as variáveis. Uma revisão da tabulação cruzada da Tabela 2.11 revela que os preços de refeições mais elevados estão associados com os restaurantes de qualidade mais elevada, e que os preços de refeições mais baixos estão associados com os restaurantes de qualidade mais baixa.

O ato de transformarmos em porcentagens de linha ou em porcentagens de coluna os lançamentos efetuados em uma tabulação cruzada pode fornecer-nos mais *insight* a respeito da relação entre as duas variáveis. Quanto às porcentagens de linha, os resultados de dividirmos cada freqüência apresentada na Tabela 2.11 pelo total de sua respectiva linha são mostrados na Tabela 2.12. Cada linha dessa última tabela refere-se a uma distribuição de freqüência percentual do preço das refeições correspondente a uma das categorias de avaliação da qualidade. Da análise dos restaurantes com avaliação da qualidade mais baixa (bom), observamos que as maiores porcentagens referem-se aos restaurantes mais baratos (50% têm preços de refeições de US$ 10-19 e 47,6%, de US$ 20-29). Dos restaurantes com avaliação da qualidade mais alta (excelente), observamos que as maiores porcentagens referem-se aos restaurantes mais caros (42,4% têm preços de refeições de US$ 30-39, e 33,4% têm preços de refeições de US$ 40-49). Desse modo, continuamos a notar que as refeições mais caras estão associadas aos restaurantes com qualidade mais elevada.

As tabulações cruzadas são amplamente usadas quando se quer examinar a relação entre duas variáveis. Na prática, os relatórios finais de muitos estudos estatísticos incluem um grande número de tabelas de tabulação cruzada. Na pesquisa dos restaurantes de Los Angeles, a tabulação cruzada baseia-se em uma variável qualitativa (avaliação da qualidade) e em uma variável quantitativa (preço das refeições). Tabulações cruzadas também podem ser desenvolvidas tanto quando ambas as variáveis são qualitativas

como quando são quantitativas. Entretanto, quando são usadas variáveis quantitativas, devemos primeiramente criar classes para os valores da variável. Assim, no exemplo dos restaurantes agrupamos os preços das refeições em quatro classes (US$ 10-19, US$ 20-29, US$ 30-39 e US$ 40-49).

Tabela 2.12 Porcentagens de linha para cada categoria de avaliação da qualidade

	Preço da Refeição				
Avaliação da Qualidade	$10–19	$20–29	$30–39	$40–49	Total
Bom	50,0	47,6	2,4	0,0	100
Ótimo	22,7	42,7	30,6	4,0	100
Excelente	3,0	21,2	42,4	33,4	100

O Paradoxo de Simpson

Freqüentemente, os dados de duas ou mais tabulações cruzadas são combinados ou agregados para produzir uma tabulação cruzada resumida que mostre como as duas variáveis estão relacionadas. Nesses casos, devemos ser cuidadosos ao tirar conclusões a respeito da relação entre as duas variáveis da tabulação cruzada agregada. Em alguns casos, a conclusão baseada na tabulação cruzada agregada pode ser completamente invertida se olharmos para os dados não-agregados, uma ocorrência que é conhecida como **paradoxo de Simpson**. A fim de oferecermos uma ilustração do paradoxo de Simpson, vamos considerar um exemplo envolvendo a análise de um veredito dado por dois juízes em dois tipos de tribunais.

Os juízes Ron Luckett e Dennis Kendall presidiram os julgamentos na Common Pleas Court[6] e na Municipal Court (Tribunal Municipal) durante os últimos três anos. Alguns dos veredictos que eles proferiram sofreram apelação. Na maioria desses casos, os tribunais de apelação confirmaram os veredictos originais, mas, em alguns casos, esses veredictos foram revertidos. Foi desenvolvida uma tabulação cruzada correspondente a cada juiz, tendo como base duas variáveis: Veredicto (confirmado ou revertido) e Tipo de Tribunal (Common Pleas e Municipal). Suponha que as duas tabulações cruzadas tenham sido então combinadas agregando-se os dados sobre o tipo de tribunal. A tabulação cruzada agregada resultante conteria duas variáveis: Veredicto (confirmado ou revertido) e Juiz (Luckett ou Kendall). Essa tabulação cruzada mostra o número de apelações em que o veredicto foi confirmado e o número em que o veredicto foi revertido para ambos os juízes. A tabulação cruzada a seguir mostra esses resultados juntamente com as porcentagens de coluna entre parênteses com cada valor.

	Juiz		
Veredicto	Luckett	Kendall	Total
Confirmado	129 (86%)	110 (88%)	239
Revertido	21 (14%)	15 (12%)	36
Total (%)	150 (100%)	125 (100%)	275

Uma revisão das porcentagens de coluna mostra que 14% dos veredictos do juiz Luckett foram revertidos, mas somente 12% dos veredictos do juiz Kendall foram revertidos. Desse modo, poderíamos concluir que o juiz Kendall realiza um trabalho melhor, porque uma porcentagem maior dos seus veredictos é confirmada. Entretanto, surge um problema com essa conclusão.

As tabulações cruzadas a seguir mostram os casos julgados pelos juízes Luckett e Kendall nos dois tribunais; as porcentagens de coluna também são indicadas entre parênteses com cada valor.

	Juiz Luckett				Juiz Kendall		
Veredicto	Apelações Comuns	Corte Municipal	Total	Veredicto	Apelações Comuns	Corte Municipal	Total
Confirmado	29 (91%)	100 (85%)	129	Confirmado	90 (90%)	20 (80%)	110
Revertido	3 (9%)	18 (15%)	21	Revertido	10 (10%)	5 (20%)	15
Total (%)	32 (100%)	118 (100%)	150	Total (%)	100 (100%)	25 (100%)	125

[6] NT: *Common Pleas Court*: 1. Tribunal de Primeira Instância. Em alguns estados norte-americanos, um tribunal que tem jurisdição geral e original sobre questões civis e criminais. 2. Na Inglaterra, antigo tribunal superior com jurisdição sobre questões civis (direito).

Da tabulação cruzada e das porcentagens de coluna referentes ao juiz Luckett, notamos que seus veredictos foram confirmados em 91% nos casos da Corte de Apelação Comum e em 85% dos casos da Corte Municipal. Da tabulação cruzada e das porcentagens de coluna referentes ao juiz Kendall, notamos que seus veredictos foram confirmados em 90% dos casos da Corte de Apelação Comum e em 80% dos casos da Corte Municipal. Comparando as porcentagens de coluna correspondentes aos dois juízes, notamos que o juiz Luckett demonstra uma atuação melhor que a do juiz Kendall em ambos os tribunais. Esse resultado contradiz a conclusão a que chegamos quando agregamos os dados entre os dois tribunais para a tabulação cruzada original. Parecia então que o juiz Kendall tinha o melhor desempenho. Esse exemplo ilustra o paradoxo de Simpson.

A tabulação cruzada original foi obtida agregando-se os dados contidos nas tabulações cruzadas separadas referentes aos dois tribunais. Note que, para ambos os juízes, a porcentagem de apelações que resultaram em reversões foi muito mais elevada na Corte Municipal do que na Corte de Apelação Comum. Uma vez que o juiz Luckett julgou uma porcentagem maior de seus casos na Corte Municipal, os dados agregados favoreciam o juiz Kendall. No entanto, quando olhamos para as tabulações cruzadas correspondentes aos dois tribunais separadamente, o juiz Luckett mostra claramente o melhor desempenho. Dessa forma, considerando a tabulação cruzada original, vemos que o *tipo de tribunal* é uma variável oculta que não pode ser ignorada quando se avalia o desempenho dos dois juízes.

Em virtude do paradoxo de Simpson, precisamos ser especialmente cuidadosos ao tirar conclusões utilizando dados agregados. Antes de tirar conclusões a respeito da relação entre duas variáveis mostradas por uma tabulação cruzada envolvendo dados agregados, você deve investigar se alguma variável oculta poderia afetar os resultados.

Diagramas de Dispersão e Linha de Tendência

Um **diagrama de dispersão** é uma apresentação gráfica da relação existente entre duas variáveis, e uma **linha de tendência** é uma linha que fornece uma aproximação da relação. Como ilustração, considere a relação publicidade/vendas de uma loja de equipamentos de som em São Francisco. Em dez ocasiões durante os três últimos meses, a loja utilizou comerciais de televisão de fins de semana para promover as vendas em suas lojas. Os gerentes querem verificar se existe uma relação entre o número de comerciais exibidos e as vendas na loja durante a semana seguinte. Dados de amostra correspondentes às dez semanas, com as vendas expressas em centenas de dólares, são mostrados na Tabela 2.13. A Figura 2.7 apresenta o diagrama de dispersão e a linha de tendência[7] dos dados da Tabela 2.13. O número de comerciais (x) é indicado no eixo horizontal, e as vendas (y) são mostradas no eixo vertical. Para a semana 1, $x = 2$ e $y = 50$. Um ponto com essas coordenadas é assinalado no diagrama de dispersão. Pontos idênticos são assinalados para as outras nove semanas. Note que durante duas das semanas foi exibido um comercial, durante duas das semanas foram exibidos dois comerciais e assim por diante.

O diagrama de dispersão completo da Figura 2.7 indica uma possível relação entre o número de comerciais e as vendas. Um maior número de vendas está associado a um maior número de comerciais. A relação não é perfeita em termos de que todos os pontos não estão em uma linha reta. Entretanto, o padrão geral dos pontos e a linha de tendência sugerem que a relação global é positiva.

ARQUIVO DA INTERNET
Stereo

Tabela 2.13 Dados de amostra da loja de equipamentos de som

Semana	Número de Comerciais x	Vendas (\$100s) y
1	2	50
2	5	57
3	1	41
4	3	54
5	4	54
6	1	38
7	5	63
8	3	48
9	4	59
10	2	46

[7] A equação da linha de tendência é $y = 4{,}95x + 36{,}15$. O declive da linha de tendência é 4,95, e o ponto de interseção com y (o ponto em que a linha intercepta o eixo y) é 36,15. Discutiremos detalhadamente a interpretação do declive e do ponto de interseção com y da linha de tendência linear no Capítulo 12, quando estudaremos as regressões lineares simples.

Alguns padrões gerais dos diagramas de dispersão e os tipos de relação que eles sugerem são expostos na Figura 2.8. O painel superior esquerdo descreve uma relação positiva similar à do exemplo do número de comerciais e vendas.

Figura 2.7 Diagrama de dispersão e linha de tendência da loja de equipamentos de som

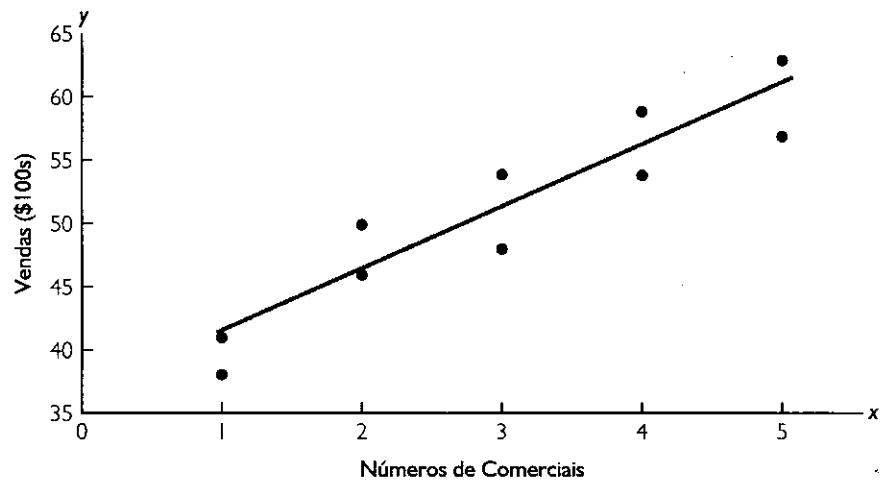
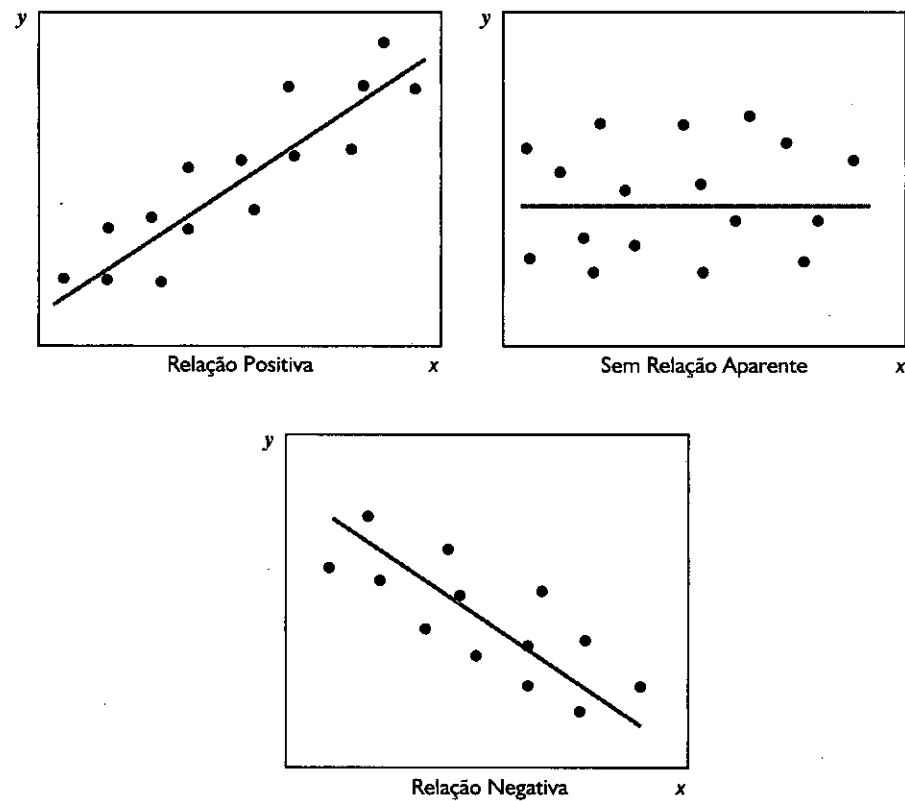



Figura 2.8 Tipos de relação representados por diagramas de dispersão

No painel superior direito, o diagrama de dispersão não mostra nenhuma relação aparente entre as variáveis. O painel inferior representa uma relação negativa, em que y tende a decrescer à medida que x aumenta.

Exercícios

Métodos

AUTOTESTE

ARQUIVO DA INTERNET
Crosstab

29. Os dados a seguir referem-se a 30 observações envolvendo duas variáveis qualitativas, x e y. As categorias correspondentes a x são A, B e C; as categorias correspondentes a y são 1 e 2.

Observação	x	y	Observação	x	y
1	A	1	16	B	2
2	B	1	17	C	1
3	B	1	18	B	1
4	C	2	19	C	1
5	B	1	20	B	1
6	C	2	21	C	2
7	B	1	22	B	1
8	C	2	23	C	2
9	A	1	24	A	1
10	B	1	25	B	1
11	A	1	26	C	2
12	B	1	27	C	2
13	C	2	28	A	1
14	C	2	29	B	1
15	C	2	30	B	2

a. Desenvolva uma tabulação cruzada dos dados, sendo x a variável linha e y a variável coluna.
b. Calcule as porcentagens de linhas.
c. Calcule as porcentagens de colunas.
d. Qual é a relação, se houver, entre x e y?

AUTOTESTE

ARQUIVO DA INTERNET
Scatter

30. As 20 observações seguintes referem-se a duas variáveis quantitativas, x e y.

Observação	x	y	Observação	x	y
1	−22	22	11	−37	48
2	−33	49	12	34	−29
3	2	8	13	9	−18
4	29	−16	14	−33	31
5	−13	10	15	20	−16
6	21	−28	16	−3	14
7	−13	27	17	−15	18
8	−23	35	18	12	17
9	14	−5	19	−20	−11
10	3	−3	20	−7	−22

a. Desenvolva um diagrama de dispersão para a relação entre x e y.
b. Qual é a relação, se houver, entre x e y?

Aplicações

31. A tabulação cruzada apresentada a seguir mostra a renda familiar por nível educacional de chefes de família (*Statistical Abstract of the United States: 2002*).

	Renda Familiar ($1.000s)					
Nível Educacional	Abaixo de 25	25,0–49,9	50,0–74,9	75,0–99,9	100 ou mais	Total
Sem diploma do ensino médio	9.285	4.093	1.589	541	354	15.862
Com diploma do ensino médio	10.150	9.821	6.050	2.737	2.028	30.786
Superior incompleto	6.011	8.221	5.813	3.215	3.120	26.380
Superior completo	2.138	3.985	3.952	2.698	4.748	17.521
Pós-graduação	8.13	1.497	1.815	1.589	3.765	9.479
Total	28.397	27.617	19.219	10.780	14.015	100.028

a. Calcule as porcentagens de linha e identifique as distribuições de freqüência percentual de renda das famílias em que o chefe da casa tem diploma do ensino médio e das famílias em que o chefe da casa tem o grau de bacharel.
b. Qual porcentagem de famílias chefiadas por pessoas com diploma do ensino médio ganha US$ 75.000 ou mais? Qual porcentagem de famílias chefiadas por pessoas que receberam graus de bacharel ganham US$ 75.000 ou mais?
c. Construa histogramas de freqüência percentual da renda de família chefiada por pessoas com diploma do ensino médio e daquelas chefiadas por pessoas com grau de bacharel? Há alguma relação clara entre a renda familiar e o nível educacional?

32. Consulte novamente a tabulação cruzada de renda familiar por nível educacional mostrada no Exercício 31.
a. Calcule as porcentagens de coluna e identifique as distribuições de freqüência percentual apresentadas. Qual porcentagem dos chefes de família não têm diploma do ensino médio?
b. Qual porcentagem das famílias que ganham US$ 100.000 ou mais era chefiada por pessoas que têm escolaridade superior ao grau de bacharel? Qual porcentagem das famílias chefiadas por pessoas com escolaridade superior ao grau de bacharel ganharam mais de US$ 100.000? Por que essas duas porcentagens são diferentes?
c. Compare as distribuições de freqüência percentual relativas às famílias que ganham "menos de 25", "100 ou mais" e o "total". Comente a relação entre a renda familiar e o nível educacional do chefe de família.

33. Recentemente, a gerência do Oak Tree Golf Course recebeu algumas reclamações sobre o estado dos *greens*[8] nos campos de golfe. Vários jogadores queixaram-se de que os *greens* eram muito rápidos. Em vez de reagir aos comentários de apenas alguns jogadores, a Golf Association realizou uma pesquisa de 100 golfistas masculinos e femininos. Os resultados da pesquisa estão resumidos a seguir:

Golfistas Masculinos			Golfistas Femininos		
	Estado dos *greens*			Estado dos *greens*	
Handicaps[9] (Desvantagem)	Muito Rápidos	Ótimo	*Handicaps*[9] (Desvantagem)	Muito Rápidos	Ótimo
Menos de 15	10	40	Menos de 15	1	9
15 ou mais	25	25	15 ou mais	39	51

a. Combine essas duas tabulações cruzadas em uma que contenha as palavras masculino e feminino como rótulos de linha e as palavras muito rápidos e ótimo como rótulos de coluna. Qual grupo mostra a maior porcentagem de pessoas que dizem que os *greens* são muito rápidos?
b. Consulte as tabulações cruzadas iniciais. Dos jogadores com menos *handicap* (melhores jogadores), qual grupo (masculino ou feminino) exibe a maior porcentagem dos que dizem que os *greens* são muito rápidos?
c. Consulte as tabulações cruzadas iniciais. Dos jogadores com maiores *handicaps*, qual grupo (masculino ou feminino) exibe a maior porcentagem dos que dizem que os *greens* são muito rápidos?
d. Quais conclusões você seria capaz de tirar a respeito das preferências de homens e mulheres no que se refere à velocidade dos *greens*? As conclusões que você tira da parte (a) são consistentes quando comparadas com as partes (b) e (c)? Explique quaisquer inconsistências claras.

34. A Tabela 2.14 apresenta dados financeiros de uma amostra de 36 empresas cujos títulos são negociados na Bolsa de Valores de Nova York (*Investor's Business Daily*, 7 de abril de 2000). Os dados sobre Vendas/Margens de Lucro/RPL são um compósito baseado na taxa de crescimento das vendas da empresa, suas margens de lucro e seu retorno sobre o patrimônio líquido (RPL).

[8] NT: *Greens* – A área coberta de relva cuidadosamente tratada ao redor de cada buraco para facilitar a tacada (Golfe).
[9] NT: *Handicap* – O golfe tem um sistema denominado *handicap* que possibilita a jogadores de diferentes níveis de habilidade disputarem uma partida entre si. O *handicap* possibilita tacadas de "vantagem" ao jogador menos experiente, as quais devem ser descontadas ao final do jogo. Jogadores profissionais jogam com *handicap* 0 (Golfe).

Tabela 2.14 Dados financeiros de uma amostra de 36 empresas

Empresa	Lucro por Ação (LPA)	Força Relativa de Preços	Força Relativa do Setor	Vendas/Margens de Lucro/Retorno sobre o Patrimônio Líquido (RLP)
Advo	81	74	B	A
Alaska Air Group	58	17	C	B
Alliant Tech	84	22	B	B
Atmos Energy	21	9	C	E
Bank of Am.	87	38	C	A
Bowater PLC	14	46	C	D
Callaway Golf	46	62	B	E
Central Parking	76	18	B	C
Dean Foods	84	7	B	C
Dole Food	70	54	E	C
Elec. Data Sys.	72	69	A	B
Fed. Dept. Store	79	21	D	B
Gateway	82	68	A	A
Goodyear	21	9	E	D
Hanson PLC	57	32	B	B
ICN Pharm.	76	56	A	D
Jefferson Plt.	80	38	D	C
Kroger	84	24	D	A
Mattel	18	20	E	D
McDermott	6	6	A	C
Monaco	97	21	D	A
Murphy Oil	80	62	B	B
Nordstrom	58	57	B	C
NYMAGIC	17	45	D	D
Office Depot	58	40	B	B
Payless Shoes	76	59	B	B
Praxair	62	32	C	B
Reebok	31	72	C	E
Safeway	91	61	D	A
Teco Energy	49	48	D	B
Texaco	80	31	D	C
US West	60	65	B	A
United Rental	98	12	C	A
Wachovia	69	36	E	B
Winnebago	83	49	D	A
York International	28	14	D	B

Fonte: *Investor's Business Daily*, 7 de abril de 2000.

a. Prepare uma tabulação cruzada dos dados sobre vendas/margens de lucro/RPL (linhas) e lucro por ação (colunas). Use as classes 0-19, 20-39, 40-59, 60-79 e 80-99 para o lucro por aAção.

b. Calcule as porcentagens de linha e comente a possível relação entre as variáveis.

35. Consulte os dados da Tabela 2.14.

a. Prepare uma tabulação cruzada dos dados sobre vendas/margens de lucro/RPL e força relativa do setor.

b. Prepare uma distribuição de freqüência dos dados sobre vendas/margens de lucro/RPL.

c. Prepare uma distribuição de freqüência dos dados sobre a força relativa do setor.

d. Como a tabulação cruzada ajudou a preparar as distribuições de freqüência nas partes (b) e (c)?

36. Consulte os dados da Tabela 2.14.

a. Prepare um diagrama de dispersão dos dados sobre o lucro por ação e força relativa de preços.

b. Comente a relação, se houver, entre as variáveis. (O significado da avaliação do lucro por ação é descrito no Exercício 34. A força relativa de preços é uma medida da variação no preço das ações ao longo dos últimos 12 meses. Valores mais elevados indicam maior força.)

37. A National Football League avalia os calouros posição por posição em uma escala que varia de 5 a 9. As avaliações são interpretadas da seguinte maneira: 8-9 devem começar primeiro ano; 7,0-7,9 devem começar; 6,0-6,9 formarão o time reserva e 5,0-5-9 poderão integrar o clube e contribuir, quando

necessário. A Tabela 2.15 mostra a posição, peso, velocidade (segundos para percorrer 36,5 m) e as classificações de 40 candidados à NFL (*USA Today*, 14 de abril de 2000).

a. Prepare uma tabulação cruzada dos dados sobre posição (linhas) e velocidade (colunas). Use classes de 4,00-4,49; 4,50-4,99; 5,00-5,49; e 5,50-5,99 para a velocidade.
b. Comente a relação entre posição e velocidade baseando-se na tabulação cruzada desenvolvida no item (a).
c. Desenvolva um diagrama de dispersão dos dados sobre velocidade e avaliação. Use o eixo vertical para avaliação.
d. Comente a relação, se houver, entre velocidade e avaliação.

Tabela 2.15 Avaliações da National Football League de 40 candidatos ao *draft*[10]

Observação	Nome	Posição	Peso (kg)	Velocidade	Avaliação
1	Peter Warrick	*Wide receiver*[11]	87,99	4,53	9
2	Plaxico Buress	*Wide receiver*	104,8	4,52	8,8
3	Sylvester Morris	*Wide receiver*	97,97	4,59	8,3
4	Travis Taylor	*Wide receiver*	90,26	4,36	8,1
5	Laveranues Coles	*Wide receiver*	87,09	4,29	8
6	Dez White	*Wide receiver*	98,88	4,49	7,9
7	Jerry Porter	*Wide receiver*	100,24	4,55	7,4
8	Ron Dugans	*Wide receiver*	93,44	4,47	7,1
9	Todd Pinkston	*Wide receiver*	76,66	4,37	7
10	Dennis Northcutt	*Wide receiver*	79,38	4,43	7
11	Anthony Lucas	*Wide receiver*	87,99	4,51	6,9
12	Darrell Jackson	*Wide receiver*	89,36	4,56	6,6
13	Danny Farmer	*Wide receiver*	98,43	4,6	6,5
14	Sherrod Gideon	*Wide receiver*	78,47	4,57	6,4
15	Trevor Gaylor	*Wide receiver*	90,26	4,57	6,2
16	Cosey Coleman	*Guard*[12]	146,05	5,38	7,4
17	Travis Claridge	*Guard*	137,44	5,18	7
18	Kaulana Noa	*Guard*	143,79	5,34	6,8
19	Leander Jordan	*Guard*	149,68	5,46	6,7
20	Chad Clifton	*Guard*	151,45	5,18	6,3
21	Manula Savea	*Guard*	139,71	5,32	6,1
22	Ryan Johanningmeir	*Guard*	140,61	5,28	6
23	Mark Tauscher	*Guard*	144,24	5,37	6
24	Blaine Saipaia	*Guard*	145,60	5,25	6
25	Richard Mercier	*Guard*	133,80	5,34	5,8
26	Damion McIntosh	*Guard*	148,78	5,31	5,3
27	Jeno James	*Guard*	145,15	5,64	5
28	Al Jackson	*Guard*	137,89	5,2	5
29	Chris Samuels	*Offensive tackle*[13]	147,41	4,95	8,5
30	Stockar McDougle	*Offensive tackle*	163,74	5,5	8
31	Chris McIngosh	*Offensive tackle*	142,88	5,39	7,8
32	Adrian Klemm	*Offensive tackle*	139,25	4,98	7,6
33	Todd Wade	*Offensive tackle*	147,87	5,2	7,3
34	Marvel Smith	*Offensive tackle*	145,15	5,36	7,1
35	Michael Thompson	*Offensive tackle*	130,18	5,05	6,8
36	Bobby Williams	*Offensive tackle*	150,9	5,26	6,8
37	Darnell Alford	*Offensive tackle*	151,5	5,55	6,4
38	Terrance Beadles	*Offensive tackle*	141,2	5,15	6,3
39	Tutan Reyes	*Offensive tackle*	135,2	5,35	6,1
40	Greg Robinson-Ran	*Offensive tackle*	151,4	5,59	6

ARQUIVO DA INTERNET
NFL

[10] NT: *Draft* – Seleção de jovens atletas na National Football League (futebol americano).

[11] NT: *Wide receiver* – Jogador que recebe os lançamentos em linha avançada para conseguir o máximo de jardas à frente (futebol americano).

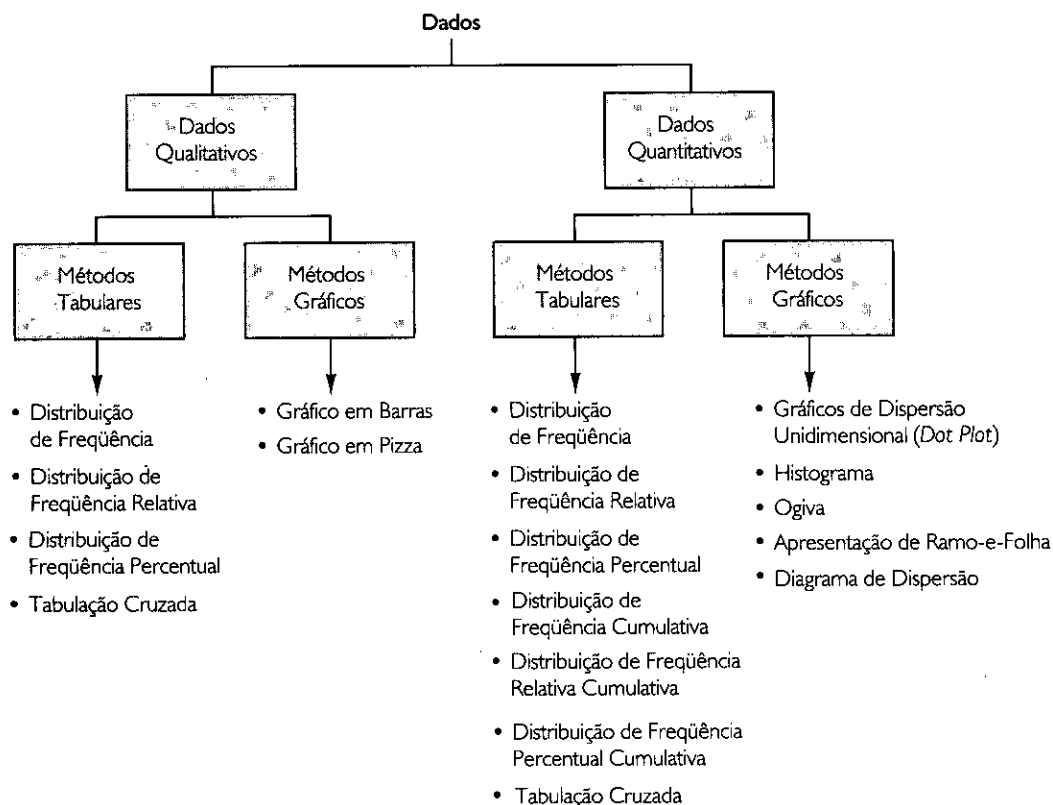
[12] NT: *Guard* – Um atleta da linha ofensiva (futebol americano).

[13] NT: *Offensive tackle* – Jogador integrante da primeira linha de ataque, a qual é composta pelos maiores jogadores do time, sendo a sua função bloquear a defesa adversária (futebol americano).

Resumo

Um conjunto de dados, mesmo que moderadamente pequeno, com freqüência é difícil de ser interpretado diretamente na forma em que é coletado. Os métodos tabulares e gráficos fornecem meios de se organizar e sintetizar dados de modo que certos padrões sejam revelados e os dados sejam mais facilmente interpretados. Distribuições de freqüência, distribuições de freqüência relativa, distribuições de freqüência percentual, gráficos em barras e gráficos em setores ("pizza") foram indicados como procedimentos tabulares e gráficos para sintetizar dados qualitativos. Distribuições de freqüência, distribuições de freqüência relativa, distribuições de freqüência percentual, histogramas, distribuições de freqüência cumulativa, distribuições de freqüência relativa cumulativa, distribuições de freqüência percentual cumulativa e ogivas foram apresentadas como meios de sintetizar dados quantitativos. Uma apresentação de ramo-e-folha constitui uma técnica de análise exploratória de dados que pode ser usada para sintetizar dados referentes a duas variáveis. O diagrama de dispersão foi exposto como um método gráfico para exibir a relação entre duas variáveis quantitativas. A Figura 2.9 mostra os métodos tabular e gráfico apresentados neste capítulo.

Quando se trata de grandes conjuntos de dados, softwares de computador são fundamentais para se construir sumários tabulares e gráficos de dados. Nos dois apêndices deste capítulo mostraremos como o Minitab e o Excel podem ser usados com essa finalidade.

Figura 2.9 Métodos tabulares e gráficos para sintetizar dados

Glossário

Dados qualitativos Rótulos ou nomes utilizados para identificar categorias de itens semelhantes.

Dados quantitativos Valores numéricos que indicam quantidade.

Distribuição de freqüência Sumário tabular dos dados que mostra a fração ou proporção dos valores de dados em cada uma das diversas classes não sobrepostas.

Distribuição de freqüência relativa Sumário tabular dos dados que mostra a fração ou proporção dos valores de dados em cada uma das diversas classes não sobrepostas.

Distribuição de freqüência percentual Sumário tabular dos dados que mostra a porcentagem de valores de dados em cada uma das diversas classes não sobrepostas.

Gráfico em barras Dispositivo gráfico para representar dados qualitativos que foram sintetizados em uma distribuição de freqüência, de freqüência relativa ou de freqüência percentual.

Gráfico em setores ("Pizza") Dispositivo gráfico para apresentar sumários de dados, baseado na subdivisão de um círculo em setores que correspondem à freqüência relativa de cada classe.

Ponto médio da classe O valor intermediário entre os limites de classe superior e inferior.

Gráficos de dispersão unidimensional (*dot plot*) Dispositivo gráfico que sintetiza dados por meio do número de pontos acima de cada valor no eixo horizontal.

Histograma Representação gráfica de uma distribuição de freqüência, de uma distribuição de freqüência relativa ou de uma distribuição de freqüência percentual de dados quantitativos, a qual é construída colocando-se os intervalos de classe no eixo horizontal e as freqüências, freqüências relativas ou freqüências percentuais no eixo vertical.

Distribuição de freqüência cumulativa Sumário tabular de dados quantitativos que mostra o número de valores de dados menores ou iguais ao limite superior de classe de cada uma das classes.

Distribuição de freqüência relativa cumulativa Sumário tabular de dados quantitativos que mostra a fração ou proporção dos valores de dados que são menores ou iguais ao limite superior de classe de cada uma das classes.

Distribuição de freqüência percentual cumulativa Sumário tabular de dados quantitativos que mostra a porcentagem dos valores de dados que são menores ou iguais ao limite superior de cada uma das classes.

Ogiva Gráfico de uma distribuição cumulativa.

Análise exploratória de dados Métodos que utilizam cálculos aritméticos simples e gráficos fáceis de desenhar para sintetizar dados rapidamente.

Apresentação de ramo-e-folha Técnica de análise exploratória de dados que simultaneamente classifica pela ordem os dados quantitativos e fornece *insight* sobre a forma da distribuição.

Tabulação cruzada Sumário tabular dos dados correspondentes a duas variáveis. As classes de uma variável são representadas pelas linhas; as classes da outra variável são representadas pelas colunas.

Paradoxo de Simpson Conclusões tiradas a partir de duas ou mais tabulações cruzadas que podem ser invertidas quando os dados são agregados em uma única tabulação cruzada.

Diagrama de dispersão Representação gráfica da relação entre duas variáveis quantitativas. Uma variável é mostrada no eixo horizontal e a outra variável, no eixo vertical.

Linha de tendência Linha que fornece uma aproximação da relação entre duas variáveis.

Fórmulas-Chave

Freqüência Relativa

$$\frac{\text{Freqüência da classe}}{n} \qquad (2.1)$$

Amplitude aproximada de classe

$$\frac{\text{Maior valor dos dados} - \text{Menor valor dos dados}}{\text{Número de classes}} \qquad (2.2)$$

Exercícios Suplementares

38. Os cinco veículos mais vendidos em 2003 (nos Estados Unidos) foram a picape Chevrolet Silverado/C/K, a picape Dodge Ram, a picape Ford F-Series, o Honda Accord e o Toyota Camry (*Motor Trend*, 2003). Dados de uma amostra de 50 compras de veículos são apresentados na Tabela 2.16.

ARQUIVO DA INTERNET
AutoData

Tabela 2.16 Dados de 50 compras de veículos

Silverado	Ram	Accord	Camry	Camry
Silverado	Silverado	Camry	Ram	F-Series
Ram	F-Series	Accord	Ram	Ram
Silverado	F-Series	F-Series	Silverado	Ram
Ram	Ram	Accord	Silverado	Camry
F-Series	Ram	Silverado	Accord	Silverado
Camry	F-Series	F-Series	F-Series	Silverado
F-Series	Silverado	F-Series	F-Series	Ram
Silverado	Silverado	Camry	Camry	F-Series
Silverado	F-Series	F-Series	Accord	Accord

a. Desenvolva uma distribuição de freqüência e de freqüência percentual.
b. Qual é a caminhonete mais vendida e qual é o carro de passageiros mais vendido?
c. Apresente um gráfico em setores ("pizza").

39. Cada uma das empresas da *Fortune 1000* pertence a um dos diversos setores industriais (*Fortune*, 17 de abril de 2000). Segue-se uma amostra de 20 empresas, com suas classificações por setor.

Empresa	Classificação por Setor	Empresa	Classificação por Setor
IBP	Alimentos	Borden	Alimentos
Intel	Produtos eletrônicos	McDonnell Douglas	Indústria aeroespacial
Coca-Cola	Bebidas	Morton International	Produtos químicos
Union Carbide	Produtos químicos	Quaker Oats	Alimentos
General Electric	Produtos eletrônicos	PepsiCo	Bebidas
Motorola	Produtos eletrônicos	Maytag	Produtos eletrônicos
Kellog	Alimentos	Textron	Indústria aeroespacial
Dow Chemical	Produtos químicos	Sara Lee	Alimentos
Campbell Soup	Alimentos	Harris	Produtos eletrônicos
Ralston Purina	Alimentos	Eaton	Produtos eletrônicos

a. Apresente uma distribuição de freqüência mostrando o número de empresas de cada indústria.
b. Apresente uma distribuição de freqüência percentual.
c. Apresente um gráfico em barras dos dados.

40. Foi feita a seguinte pergunta aos Top 100 Teachers da *Golf Magazine*: "Qual é a área mais crítica que impede os golfistas de atingirem seu pleno potencial?" As respostas possíveis foram a falta de precisão, tacadas de aproximação à bandeira (*approach*) malfeitas, fraca abordagem mental, falta de força física, falta de prática, *putting*[14] ruim, jogadas curtas malfeitas e decisões estratégicas ruins. Os dados obtidos foram os seguintes (*Golf Magazine*, fevereiro de 2002):

ARQUIVO DA INTERNET
Golf

Abordagem mental	Abordagem mental	Jogada curta	Jogada curta	Jogada curta
Prática	Precisão	Abordagem mental	Precisão	*Putting*
Força	Tacadas de aproximação	Precisão	Jogada curta	*Putting*
Precisão	Abordagem mental	Abordagem mental	Precisão	Força
Precisão	Precisão	Jogada curta	Força	Jogada curta
Precisão	*Putting*	Abordagem mental	Decisões estratégicas	Precisão
Jogada curta	Força	Abordagem mental	Abordagem mental	Jogada curta
Prática	Prática	Abordagem mental	Força	Força
Abordagem mental	Jogada curta	Abordagem mental	Jogada curta	Decisões estratégicas
Precisão	Jogada curta	Precisão	Abordagem mental	Jogada curta
Abordagem mental	*Putting*	Abordagem mental	Abordagem mental	*Putting*
Prática	*Putting*	Prática	Jogada curta	*Putting*

[14] NT: *Putting* – *Tacada* de curto alcance (golfe).

Força	Abordagem mental	Jogada curta	Prática	Decisões estratégicas
Precisão	Jogada curta	Precisão	Prática	*Putting*
Precisão	Jogada curta	Precisão	Jogada curta	*Putting*
Precisão	Tacadas de aproximação	Jogada curta	Abordagem mental	Prática
Jogada curta	Jogada curta	Decisões estratégicas	Jogada curta	Jogada curta
Prática	Prática	Jogada curta	Prática	Decisões estratégicas
Abordagem mental	Decisões estratégicas	Decisões estratégicas	Força	Jogada curta
Precisão	Prática	Prática	Prática	Precisão

a. Desenvolva uma distribuição de freqüência e uma distribuição de freqüência percentual.
b. Quais áreas críticas impedem com maior freqüência que os golfistas atinjam seu pleno potencial?

Tabela 2.17 Valor nominal por ação de títulos da Dow Jones Industrial Average

Empresa	Valor Nominal por Ação	Empresa	Valor Nominal por Ação
AT&T	14,59	Home Depot	7,71
Alcoa	12,30	Honeywell	11,25
Altria Group	8,96	IBM	13,37
American Express	9,04	Intel	5,39
Boeing	12,92	International Paper	21,37
Caterpillar	16,18	Johnson & Johnson	7,79
Citigroup	15,09	J.P. Morgan Chase	20,31
Coca-Cola	4,57	McDonald's	7,30
Disney	11,28	Merck	6,89
Du Pont	14,17	Microsoft	8,49
Eastman Kodak	9,93	Procter & Gamble	8,80
ExxonMobil	10,62	SBE Communications	9,69
General Electric	5,43	3M	14,93
General Motors	35,15	United Technologies	17,36
Hewlett-Packard	7,33	Wal-Mart Stores	7,85

ARQUIVO DA INTERNET
Dow

41. Os dados da Tabela 2.17 mostram o valor nominal por ação dos 30 títulos que compõem a Dow Jones Industrial Average (*Barron's*, 10 de março de 2003).
    a. Construa uma distribuição de freqüência para sintetizar os dados. Use uma amplitude de classe 6,00.
    b. Desenvolva uma distribuição de freqüência relativa.
    c. Construa uma distribuição de freqüência cumulativa.
    d. Construa uma distribuição de freqüência relativa cumulativa.
    e. Construa um histograma como uma representação gráfica dos dados. Comente a forma da distribuição.
42. Os preços de fechamento de 40 ações ordinárias são apresentados a seguir (*Barron's*, 10 de março de 2003).

29,63	34,00	43,25	8,75	37,88	8,63	7,63	30,38	35,25	19,38
9,25	16,50	38,00	53,38	16,63	1,25	48,38	18,00	9,38	9,25
10,00	25,02	18,00	8,00	28,50	24,25	21,63	18,50	33,63	31,13
32,25	29,63	79,38	11,38	38,88	11,50	52,00	14,00	9,00	33,50

ARQUIVO DA INTERNET
Comstock

    a. Construa distribuições de freqüência e de freqüência relativa.
    b. Construa distribuições de freqüência cumulativa e de freqüência relativa cumulativa.
    c. Construa um histograma.
    d. Usando seus sumários, faça comentários e observações a respeito do preço das ações ordinárias.
43. Noventa e quatro *shadow stocks* foram registrados pela American Association of Individual Investor. O termo *shadow* indica títulos de firmas de pequeno a médio portes que não são acompanhadas de perto pelas grandes empresas corretoras. Foram fornecidas informações sobre onde o título foi negociado – Bolsa de Valores de Nova York (*New York Stock Exchange* – Nyse), American Stock Exchange (Amex) e mercado de balcão (*over-the-counter* – OTC) –, o lucro por ação e a razão preço/rendimentos da seguinte amostra de 20 *shadow stocks*.

ARQUIVO DA INTERNET
Shadon

Título	Bolsa de Valores	Lucro por Ação (US$)	Razão Preço/ Rendimentos
Chemi-Trol	OTC	0,39	27,30
Candie's	OTC	0,07	36,20
TST/Impreso	OTC	0,65	12,70
Unimed Pharm.	OTC	0,12	59,30
Skyline Chili	Amex	0,34	19,30
Cyanotech	OTC	0,22	29,30
Catalina Light.	Nyse	0,15	33,20
DDL Elect.	Nyse	0,10	10,20
Euphonix	OTC	0,09	49,70
Mesa Labs	OTC	0,37	14,40
RCM Tech.	OTC	0,47	18,60
Anuhco	Amex	0,70	11,40
Hello Direct	OTC	0,23	21,10
Hilite Industries	OTC	0,61	7,80
Alpha Tech.	OTC	0,11	34,60
Wegener Group	OTC	0,16	24,50
U.S. Home & Garden	OTC	0,24	8,70
Chalone Wine	OTC	0,27	44,40
Eng. Support Sys.	OTC	0,89	16,70
Int. Remote Imaging	Amex	0,86	4,70

a. Forneça distribuições de freqüência e de freqüência relativa dos dados das bolsas de valores. Onde os *shadow stocks* são mais arrolados?

b. Forneça distribuições de freqüência e de freqüência relativa dos dados de lucro por ação e da razão preço/rendimentos. Use as classes 0,00-19,00; 0,20-0,39 etc. para os dados de lucro por ação, e as classes 0,0-9,9, 10,0-19,9 etc. para a razão preço/rendimentos. Quais observações e comentários você pode fazer a respeito dos *shadow stocks*?

44. Uma relação da renda *per capita* organizada por estado (Estados Unidos) é apresentada a seguir (Bureau of Economic Analysis, *Current Population Survey*, março de 2000).

ARQUIVO DA INTERNET
Income

Ala.	21.500	Ky.	21.551	N.D.	21.708
Alasca	25.771	La.	21.385	Ohio	25.239
Ariz.	23.152	Maine	23.002	Okla.	21.056
Ark.	20.393	Md.	30.023	Ore.	24.775
Calif.	27.579	Mass.	32.902	Penn.	26.889
Colo.	28.821	Mich.	25.979	R.I.	26.924
Conn.	37.700	Minn.	27.667	S.C.	21.387
Del.	29.932	Miss.	18.998	S.D.	22.201
D.C.	37.325	Mo.	24.447	Tenn.	23.615
Fla.	25.922	Mont.	20.427	Texas	25.028
Ga.	25.106	Neb.	24.786	Utah	21.096
Havaí	26.210	Nev.	27.360	Vt.	24.217
Idaho	21.080	N.H.	29.219	Va.	27.489
Ill.	28.976	N.J.	33.953	Wash.	28.066
Ind.	24.302	N.M.	20.008	W. Va.	19.373
Iowa	24.007	N.Y.	31.679	Wis.	25.184
Kan.	25.049	N.C.	24.122	Wyo.	23.225

Desenvolva uma distribuição de freqüência, uma distribuição de freqüência relativa e um histograma.

45. A *Drug Store News* (setembro de 2000) forneceu dados sobre as vendas de produtos farmacêuticos das principais farmácias de venda a varejo nos Estados Unidos. Os dados a seguir referem-se a vendas anuais em milhões de dólares.

Varejista	Vendas	Varejista	Vendas
Ahold USA	$ 1.700	Medicine Shoppe	$ 1.757
CVS	12.700	Rite-Aid	8.637
Eckerd	7.739	Safeway	2.150
Kmart	1.863	Walgreens	11.660
Kroger	3.400	Wal-Mart	7.250

a. Mostre uma apresentação de ramo-e-folha.
b. Identifique os níveis anuais de vendas das menores, médias e maiores drogarias.
c. Quais são as duas maiores drogarias?

46. As temperaturas máximas e mínimas (em graus farenheit) de 20 cidades são apresentadas a seguir (*USA Today*, 9 de maio de 2000).

Cidade	Temperatura Máxima	Temperatura Mínima	Cidade	Temperatura Máxima	Temperatura Mínima
Atenas	75	54	Melbourne	66	50
Bangcoc	92	74	Montreal	64	52
Cairo	84	57	Paris	77	55
Copenhague	64	39	Rio de Janeiro	80	61
Dublin	64	46	Roma	81	54
Havana	86	68	Seul	64	50
Hong Kong	81	72	Cingapura	90	75
Johannesburgo	61	50	Sydney	68	55
Londres	73	48	Tóquio	79	59
Manila	93	75	Vancouver	57	43

ARQUIVO DA INTERNET
HighLon

a. Prepare uma apresentação de ramo-e-folha das temperaturas máximas.
b. Prepare uma apresentação de ramo-e-folha das temperaturas mínimas.
c. Compare as apresentações de ramo-e-folha dos itens (a) e (b) e faça algum comentário sobre as diferenças entre as temperaturas máximas e mínimas.
d. Use a apresentação de ramo-e-folha do item (a) para determinar o número de cidades que têm temperaturas acima de 80 graus farenheit.
e. Apresente distribuições de freqüência tanto sobre os dados relativos às temperaturas máximas como mínimas.

47. Consulte o conjunto de dados referentes às temperaturas máximas e mínimas das 20 cidades do Exercício 46.
a. Desenvolva um diagrama de dispersão para mostrar a relação entre as duas variáveis: temperatura máxima e temperatura mínima.
b. Comente a relação entre as temperaturas máxima e mínima.

48. Foi realizado um estudo a respeito da satisfação profissional de quatro ocupações. A satisfação profissional foi medida usando-se um questionário de 18 perguntas, e cada questão recebia uma pontuação de 1 a 5 para cada resposta, com as pontuações mais altas indicando maior satisfação. A soma dos pontos obtidos nas 18 questões fornece a satisfação profissional de cada indivíduo da amostra. Os dados são os seguintes:

Ocupação	Nível de Satisfação	Ocupação	Nível de Satisfação	Ocupação	Nível de Satisfação
Advogado	42	Fisioterapeuta	78	Analista de Sistemas	60
Fisioterapeuta	86	Analista de Sistemas	44	Fisioterapeuta	59
Advogado	42	Analista de Sistemas	71	Marceneiro	78
Analista de Sistemas	55	Advogado	50	Fisioterapeuta	60
Advogado	38	Advogado	48	Fisioterapeuta	50
Marceneiro	79	Marceneiro	69	Marceneiro	79
Advogado	44	Fisioterapeuta	80	Analista de Sistemas	62
Analista de Sistemas	41	Analista de Sistemas	64	Advogado	45
Fisioterapeuta	55	Fisioterapeuta	55	Marceneiro	84
Analista de Sistemas	66	Marceneiro	64	Fisioterapeuta	62
Advogado	53	Marceneiro	59	Analista de Sistemas	73
Marceneiro	65	Marceneiro	54	Marceneiro	60
Advogado	74	Analista de Sistemas	76	Advogado	60
Fisioterapeuta	52				

ARQUIVO DA INTERNET
OccupSat

a. Forneça uma tabulação cruzada da ocupação e do nível de satisfação profissional.
b. Calcule as porcentagens de linha de sua tabulação cruzada do item (a).
c. Quais observações você pode fazer a respeito do nível de satisfação profissional dessas ocupações.

49. Empresas de maior porte geram mais receita? Os dados a seguir mostram o número de empregados e a receita anual de uma amostra de 20 empresas da *Fortune 1000* (*Fortune*, 17 de abril de 2000).

ARQUIVO DA INTERNET

RevEmps

Empresa	Empregados	Receita (milhões de dólares)	Empresa	Empregados	Receita (milhões de dólares)
Sprint	77.600	19.930	American Financial	9.400	3.334
Chase Manhattan	74.801	33.710	Fluor	53.561	12.417
Computer Sciences	50.000	7.660	Phillips Petroleum	15.900	13.852
Wells Fargo	89.355	21.795	Cardinal Health	36.000	25.034
Sunbeam	12.200	2.398	Borders Group	23.500	2.999
CBS	29.000	7.510	MCI Worldcom	77.000	37.120
Time Warner	69.722	27.333	Consolidated Edison	14.269	7.491
Steelcase	16.200	2.743	IBP	45.000	14.075
Georgia-Pacific	57.000	17.796	Super Value	50.000	17.421
Toro	1.275	4.673	H&R Block	4.200	1.669

a. Prepare um diagrama de dispersão para mostrar a relação entre as variáveis Receita e Empregados.
b. Comente as possíveis relações entre as variáveis.

50. Uma pesquisa dos prédios comerciais atendidos pela Cincinnati Gas & Electric Company perguntou qual principal combustível de aquecimento era usado e em que ano o prédio fora construído. Uma tabulação cruzada parcial dos dados do levantamento é apresentada a seguir:

	Tipo de Combustível				
Ano de Construção	Eletricidade	Gás Natural	Combustível de Petróleo	Gás Propano	Outros
1973 ou antes	40	183	12	5	7
1974–1979	24	26	2	2	0
1980–1986	37	38	1	0	6
1987–1991	48	70	2	0	1

a. Conclua a tabulação cruzada mostrando os totais de linha e os totais de coluna.
b. Mostre as distribuições de freqüência correspondentes ao ano de construção e tipo de combustível.
c. Prepare uma tabulação cruzada mostrando as porcentagens de coluna.
d. Prepare uma tabulação cruzada mostrando as porcentagens de linha.
e. Comente a relação entre o ano de construção e o tipo de combustível.

51. A Tabela 2.18 contém uma parte dos dados do arquivo intitulado Fortune que se encontra no site www.thomsonlearning.com.br/estatapl.htm. Ele fornece dados sobre o patrimônio dos acionistas, valor de mercado e lucro de uma amostra de 50 empresas listadas na *Fortune 500*.

ARQUIVO DA INTERNET

Fortune

Tabela 2.18 Dados de uma amostra de 50 empresas da Fortune 500

Empresa	Patrimônio dos Acionistas (US$ 1.000)	Valor de Mercado (US$ 1.000)	Lucro (US$ 1.000)
AGCO	982,1	372,1	60,6
AMP	2698,0	12017,6	2,0
Apple Computer	1642,0	4605,0	309,0
Baxter International	2839,0	21743,0	315,0
Bergen Brunswick	629,1	2787,5	3,1
Best Buy	557,7	10376,5	94,5
Charles Schwab	1429,0	35340,6	348,5
•	•	•	•
•	•	•	•
•	•	•	•
Walgreen	2849,0	30324,7	511,0
Westvaco	2246,4	2225,6	132,0
Whirlpool	2001,0	3729,4	325,0
Xerox	5544,0	35603,7	395,0

a. Prepare uma tabulação cruzada das variáveis Patrimônio dos Acionistas e Lucro. Use as classes 0-200, 200-400, ..., 1.000-1.200 para Lucro, e as classes 0-1.200, 1.200-2.400, ..., 4.600-6.000 para o Patrimônio dos Acionistas.
b. Calcule as porcentagens de linha de sua tabulação cruzada do item (a).
c. Qual relação, se houver, você observa entre o Lucro e o Patrimônio dos Acionistas?

52. Consulte o conjunto de dados da Tabela 2.18.
a. Prepare uma tabulação cruzada das variáveis Valor de Mercado e Lucro.
b. Calcule as porcentagens de linha de sua tabulação cruzada do item (a).
c. Comente a relação, se houver, entre as variáveis.

53. Consulte o conjunto de dados da Tabela 2.18.
a. Prepare um diagrama de dispersão para mostrar a relação entre as variáveis Lucro e Patrimônio dos Acionistas.
b. Comente a relação, se houver, entre as variáveis.

54. Consulte o conjunto de dados da Tabela 2.18.
a. Prepare um diagrama de dispersão para mostrar a relação entre as variáveis Valor de Mercado e Patrimônio dos Acionistas.
b. Comente a relação, se houver, entre as variáveis.

Estudo de Caso – Pelican Stores

A Pelican Stores é uma rede de lojas de vestuário feminino que opera nos Estados Unidos. A rede realizou recentemente uma promoção na qual cupons de desconto eram enviados a clientes das lojas do ramo. Os dados coletados de uma amostra de 100 transações com cartões de crédito feitas na loja durante um dia em novembro de 2002 estão contidos no arquivo intitulado Pelican.

Tabela 2.19 Dados de uma amostra de 100 compras com cartão nas lojas Pelican

Cliente	Método de Pagamento	Artigos	Valor do Desconto	Vendas	Sexo	Estado Civil	Idade
1	Discover	1	0,00	39,50	Masculino	Casado	32
2	Proprietary Card	1	25,60	102,40	Feminino	Casada	36
3	Proprietary Card	1	0,00	22,50	Feminino	Casada	32
4	Proprietary Card	5	121,10	100,40	Feminino	Casada	28
5	Mastercard	2	0,00	54,00	Feminino	Casada	34
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.
96	Mastercard	1	0,00	39,50	Feminino	Casada	44
97	Proprietary Card	9	82,75	253,00	Feminino	Casada	30
98	Proprietary Card	10	18,00	287,59	Feminino	Casada	52
99	Proprietary Card	2	31,40	47,60	Feminino	Casada	30
100	Proprietary Card	1	11,06	28,44	Feminino	Casada	44

ARQUIVO DA INTERNET
Pelican

A Tabela 2.19 mostra uma parte do conjunto de dados. Um valor diferente de zero para a variável Desconto indica que a cliente trouxe os cupons promocionais e os usou. Para um número muito pequeno de clientes, o valor dos descontos é, de fato, maior que o valor das vendas (veja a cliente 4). O valor das vendas é líquido, sem descontos ou trocos.

A gerência das Lojas Pelican gostaria de usar esses dados de amostra para conhecer sua clientela e avaliar a promoção envolvendo cupons de desconto.

Relatório Administrativo

Use os métodos tabular e gráfico de estatística descritiva para ajudar a gerência a desenvolver um perfil dos clientes e avaliar a campanha promocional. Seu relatório deveria incluir, no mínimo, o seguinte:

1. Distribuições de freqüência percentual para as variáveis-chave.
2. Um gráfico em barras ou um gráfico em setores ("pizza") mostrando a porcentagem de compras efetuadas pelas clientes que poderiam ser atribuídas à campanha promocional.

3. Um diagrama de dispersão do tipo de clientela (normal ou atraída pela promoção) *versus* vendas. Comente quaisquer similaridades ou diferenças existentes.
4. Um diagrama de dispersão das vendas *versus* desconto relativo somente às clientes que responderam à promoção. Comente qualquer relação clara entre as vendas e os descontos.
5. Um diagrama de dispersão para explorar a relação entre as vendas e a idade das clientes.

Apêndice 2.1 – O Uso do Minitab para Apresentações Tabulares e Gráficas

O Minitab oferece extensas capacidades para a criação de sumários tabulares e gráficos de dados. Neste apêndice, mostramos como o Minitab pode ser usado para se construir diversos sumários gráficos e o sumário tabular de uma tabulação cruzada. Os métodos gráficos apresentados incluem o gráfico de dispersão unidimensional (*dot plot*), o histograma, a apresentação de ramo-e-folha e o diagrama de dispersão.

ARQUIVO DA INTERNET
Audit

Gráfico de Dispersão Unidimensional (*Dot Plot*)

Usamos os dados do tempo necessário para a conclusão das auditorias apresentados na Tabela 2.5. Os dados estão na coluna C1 de uma planilha do Minitab. As etapas a seguir gerarão um gráfico de dispersão unidimensional:

Etapa 1. Selecione o menu **Graph** e escolha **DotPlot**
Etapa 2. Selecione **One Y, Simple** e dê um clique em **OK**
Etapa 3. Quando a caixa de diálogo Dotplot-One Y aparecer:
Digite C1 na caixa **Graph Variables**
Dê um clique em **OK**

ARQUIVO DA INTERNET
Audit

Histogramas

Mostramos como construir um histograma com freqüências no eixo vertical usando os dados de tempo para conclusão das auditorias apresentados na Tabela 2.5. Os dados estão na coluna C1 de uma planilha do Minitab. Os passos a seguir gerarão um histograma dos tempos necessários para a conclusão das auditorias:

Etapa 1. Selecione o menu **Graph**
Etapa 2. Escolha **Histogram**
Etapa 3. Selecione **Simple** e dê um clique em **OK**
Etapa 4. Quando a caixa de diálogo Histogram-Simple aparecer:
Digite C1 na caixa **Graph Variables**
Dê um clique em **OK**
Etapa 5. Quando o Histograma aparecer:
Posicione o ponteiro do mouse sobre qualquer uma das barras
Dê um clique duplo
Etapa 6. Quando a caixa de diálogo Edit Bars aparecer:
Dê um clique na guia **Binning**
Selecione **Midpoint** para Interval Type
Selecione **Midpoint/Cutpoint positions** para Interval Definition
Digite 12:32/5 na caixa **Midpoint/Cutpoint positions***
Dê um clique em **OK**

ARQUIVO DA INTERNET
Aptest

Apresentação de Ramo-e-Folha

Usamos os dados do teste de aptidão apresentados na Tabela 2.9 para demonstrar a construção de uma apresentação de ramo-e-folha. Os dados estão na coluna C1 de uma planilha do Minitab. As etapas a seguir gerarão a apresentação de ramo-e-folha mostrada na Seção 2.3:

* O registro 12:35 indica que 12 é o ponto médio da primeira classe, 35 é o ponto médio da última classe e 5 é a amplitude de classe.

Etapa 1. Selecione o menu **Graph**
Etapa 2. Escolha **Stem-and-Leaf**
Etapa 3. Quando a caixa de diálogo Stem-and-Leaf aparecer:
Digite C1 na caixa **Graph Variables**
Dê um clique em **OK**

Diagrama de Dispersão

ARQUIVO DA INTERNET
Stereo

Usamos os dados da loja de equipamentos de som apresentados na Tabela 2.13 para demonstrar a construção de um diagrama de dispersão. As semanas estão numeradas de de 1 a 10 na coluna C1, os dados referentes ao número de comerciais estão na coluna C2 e os dados referentes às vendas estão na coluna C3 de uma planilha do Minitab. As etapas a seguir gerarão o diagrama de dispersão mostrado na Figura 2.7.

Etapa 1. Selecione o menu **Graph**
Etapa 2. Escolha **Scatterplot**
Etapa 3. Selecione **Simple** e dê um clique em **OK**
Etapa 4. Quando a caixa de diálogo Scatterplot-Simple aparecer:
Digite C3 sob **Y variables** e C2 sob **X variables**
Dê um clique em **OK**

Tabulação Cruzada

ARQUIVO DA INTERNET
Restaurant

Usamos os dados da Zagat's Restaurant Review, dos quais uma parte encontra-se na Tabela 2.10, para fazer nossa demonstração. Os restaurantes estão numerados de 1 a 300 na coluna C1 da planilha do Minitab. As avaliações da qualidade estão na coluna C2 e os preços das refeições estão na coluna C3.

O Minitab somente pode criar uma tabulação cruzada para variáveis qualitativas, e o preço das refeições é uma variável quantitativa. Sendo assim, precisamos primeiramente codificar os dados de preço das refeições especificando a classe à qual cada preço de refeição pertence. As etapas apresentadas a seguir codificarão os dados de preço das refeições a fim de criar quatro classes de preço de refeições na coluna C4: US$ 10–19, US$ 20–29, US$ 30–39 e US$ 40–49.

Etapa 1. Selecione o menu **Data**
Etapa 2. Escolha **Code**
Etapa 3 Escolha **Numeric to Text**
Etapa 4. Quando a caixa de diálogo Code-Numeric to Text aparecer:
Digite C3 na caixa **Code data from columns**
Digite C4 na caixa **Into columns**
Digite 10:19 na primeira caixa **Original values** e $10–19 na caixa **New** adjacente
Digite 20:29 na segunda caixa **Original values** e $20–29 na caixa **New** adjacente
Digite 30:39 na terceira caixa **Original values** e $30–39 na caixa **New** adjacente
Digite 40:49 na quarta caixa **Original values** e $40–49 na caixa **New** adjacente
Dê um clique em **OK**

Para cada preço de refeição indicado na coluna C3, a categoria de preço de refeição correspondente aparecerá agora na coluna C4. Agora, podemos desenvolver uma tabulação cruzada da avaliação da qualidade e das categorias de preço de refeição usando os dados das colunas C2 e C4. Os passos a seguir criarão uma tabulação cruzada que contém as mesmas informações mostradas na Tabela 2.11.

Etapa 1. Selecione o menu **Stat**
Etapa 2. Escolha **Tables**
Etapa 3. Escolha **Cross Tabulation and Chi-Square**
Etapa 4. Quando a caixa de diálogo Cross Tabulation and Chi-Square aparecer:
Digite C2 na caixa **For rows** e C4 na caixa **For columns**
Selecione **Counts**, abaixo da opção Display
Dê um clique em **OK**

Apêndice 2.2 – O Uso do Excel para Apresentações Tabulares e Gráficas

O Excel oferece extensas capacidades para a construção de sumários tabulares e gráficos de dados. Três das ferramentas mais potentes são a ferramenta Inserir Função, o Assistente de Gráfico e o Relatório de Tabela Dinâmica.

A Ferramenta Funções e Inserir Função

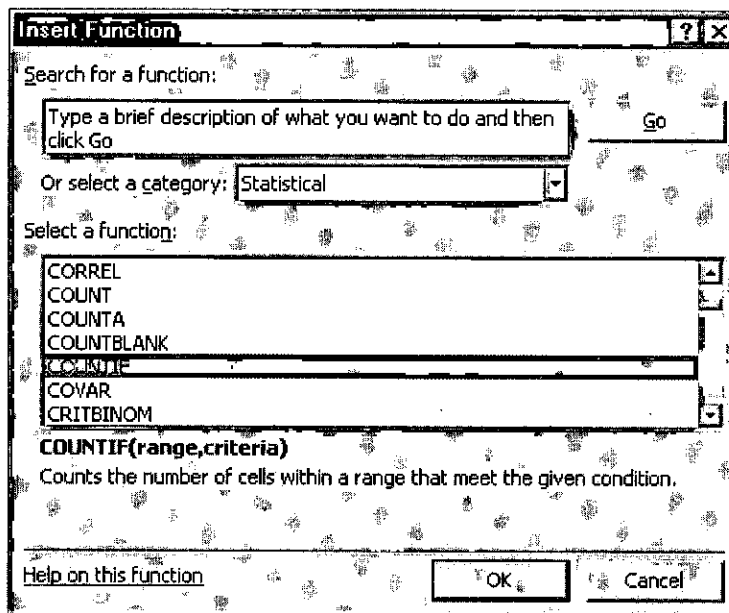
O Excel oferece uma grande variedade de funções que são úteis para a análise estatística. Se sabemos qual função queremos e a maneira de usá-la, podemos simplesmente introduzir a função diretamente em uma célula de uma planilha do Excel. Caso contrário, o Excel oferece a ferramenta Inserir Função para nos ajudar a identificar as funções disponíveis e utilizá-las.

Ferramenta Inserir Função Para acessar a ferramenta Inserir Função dê um clique na barra de fórmulas ou selecione o menu **Inserir** e escolha f_x **Função.** Aparecerá, então, a caixa de diálogo **Inserir Função** (Figura 2.10). A caixa **Ou selecione uma categoria** exibe uma relação das diversas categorias de funções do Excel; selecionamos **Estatística** na Figura 2.10. Quando a opção **Estatística** é selecionada, uma relação de todas as funções estatísticas é exibida na caixa **Selecione uma função**. Aqui, realçamos a função **CONT.SE**. Tão logo uma função é realçada, a forma apropriada da função juntamente com uma breve descrição aparece abaixo da caixa **Selecione uma função.** Para obter ajuda a respeito de como usar adequadamente a função, dê um clique em **OK.**

ARQUIVO DA INTERNET
SoftDrink

Distribuições de Freqüência Mostramos como a função **CONT.SE** pode ser usada para construirmos uma distribuição de freqüência dos dados correspondente às compras de refrigerantes apresentadas na Tabela 2.1. Consulte a Figura 2.11 à medida que descrevermos as tarefas envolvidas. A planilha de fórmulas (que mostra as funções e as fórmulas usadas) aparece em segundo plano e a planilha de valores (que mostra os resultados obtidos usando-se as funções e fórmulas) aparece em primeiro plano.

O rótulo "Marca Comprada" e os dados referentes às 50 compras de refrigerantes estão nas células A1:A51. Introduzimos também rótulos nas células C1:D1 e os nomes dos refrigerantes nas células C2:C6. A função CONT.SE do Excel pode ser usada para contar o número de vezes que cada refrigerante aparece nas células. As etapas a seguir utilizam a ferramenta Inserir Função para produzir a distribuição de freqüência que aparece no primeiro plano da Figura 2.11.

Figura 2.10 Caixa de diálogo "Inserir Função" do Excel

Etapa 1. Selecione a célula **D2**, acesse a ferramenta Inserir Função e escolha **CONT.SE** na relação de funções estatísticas

Etapa 2. Dê um clique em **OK**

Etapa 3. Quando a caixa de diálogo Argumentos da Função aparecer:

Digite $A$2:$A$51 na caixa **Intervalo**

Digite C2 na caixa **Critérios**

Dê um clique em **OK**

Etapa 4. Copie a célula D2 nas células D3:D6

A planilha de fórmulas da Figura 2.11 mostra as fórmulas de célula inseridas quando aplicamos essas etapas. A planilha de valores exibe os valores calculados através da utilização dessas fórmulas de célula; vemos que a planilha do Excel exibe a mesma distribuição de freqüência que desenvolvemos na Tabela 2.2.

Figura 2.11 A distribuição de freqüência das compras de refrigerantes construída com a função "CONT.SE" do Excel

	A	B	C	D	E
1	Marca Comprada		Refrigerante	Freqüência	
2	Coca-Cola		Coca-Cola	=COUNTIF($A$2:$A$51,C2)	
3	Coca-Cola Light		Coca-Cola Light	=COUNTIF($A$2:$A$51,C3)	
4	Pepsi-Cola		Dr. Pepper	=COUNTIF($A$2:$A$51,C4)	
5	Diet Coke		Pepsi-Cola	=COUNTIF($A$2:$A$51,C5)	
6	Coke Classic		Sprite	=COUNTIF($A$2:$A$51,C6)	
7	Coke Classic				
8	Dr. Pepper				
9	Coca-Cola Light				
10	Pepsi-Cola				
45	Pepsi-Cola				
46	Pepsi-Cola				
47	Pepsi-Cola				
48	Coca-Cola				
49	Dr. Pepper				
50	Pepsi-Cola				
51	Sprite				
52					

	A	B	C	D	E
1	Marca Comprada		Refrigerante	Freqüência	
2	Coca-Cola		Coca-Cola	19	
3	Diet Coke		Coca-Cola Light	8	
4	Pepsi-Cola		Dr. Pepper	5	
5	Diet Coke		Pepsi-Cola	13	
6	Coca-Cola		Sprite	5	
7	Coca-Cola				
8	Dr. Pepper				
9	Diet Coke				
10	Pepsi-Cola				
45	Pepsi-Cola				
46	Pepsi-Cola				
47	Pepsi-Cola				
48	Coca-Cola				
49	Dr. Pepper				
50	Pepsi-Cola				
51	Sprite				
52					

Nota: As linhas 11 a 44 estão ocultas.

Se você estiver familiarizado com a função CONT.SE e não necessitar da ajuda da ferramenta Inserir Função, pode digitar as fórmulas diretamente nas células D2:D6. Por exemplo, para contar o número de vezes que a Coca-Cola aparece, digite a seguinte fórmula na célula D2:

=CONT.SE($A$2:$A$51,C2)

Para contar o número de vezes que os outros refrigerantes aparecem, copie essa fórmula nas células D3:D6.

Muitas outras funções do Excel serão demonstradas nos apêndices dos próximos capítulos. Dependendo da complexidade da função, nós a introduziremos diretamente na célula apropriada ou utilizaremos a ferramenta Inserir Função.

Assistente de Gráfico

O Assistente de Gráfico do Excel fornece extensas capacidades para desenvolver apresentações gráficas. Essa ferramenta nos possibilita ir além daquilo que pode ser feito quando usamos somente fórmulas e funções. Mostramos como ela pode ser usada para construirmos gráficos em barras, histogramas e diagramas de dispersão.

ARQUIVO DA INTERNET

Gráficos em Barras e Histogramas Mostramos agora como o Assistente de Gráfico pode ser usado para construirmos gráficos em barras e histogramas. Vamos iniciar desenvolvendo um gráfico em barras dos dados referentes às compras de refrigerantes; construímos uma distribuição de freqüência na Figura 2.11. O gráfico que vamos desenvolver é uma extensão daquela planilha. Consulte a Figura 2.12 à medida que descrevermos as tarefas envolvidas. A planilha dos valores da Figura 2.11 aparece em segundo plano; o gráfico desenvolvido para os dados sobre refrigerantes aparece em primeiro plano.

As etapas a seguir descrevem como se pode usar o Assistente de Gráfico do Excel para construir um gráfico em barras dos dados sobre as compras de refrigerantes utilizando a distribuição de freqüência que aparece nas células C1:D6.

Etapa 1. Selecione as células C1:D6

Etapa 2. Selecione o botão **Assistente de Gráfico** na barra de ferramentas Padrão (ou selecione o menu **Inserir** e escolha a opção **Gráfico**)

Figura 2.12 Gráfico em barras das compras de refrigerantes construído com o assistente de gráfico do Excel

	A	B	C	D	E	F	G	H	I
1	**Marca Comprada**		**Refrigerante**	**Freqüência**					
2	Coca-Cola		Coca-Cola	19					
3	Coca-Cola Light		Coca-Cola Light	8					
4	Pepsi-Cola		Dr. Pepper	5					
5	Coca-Cola Light		Pepsi-Cola	13					
6	Coca-Cola		Sprite	5					
7	Coca-Cola								
8	Dr. Pepper								
9	Coca-Cola Light								
10	Pepsi-Cola								
45	Pepsi-Cola								
46	Pepsi-Cola								
47	Pepsi-Cola								
48	Coca-Cola								
49	Dr. Pepper								
50	Pepsi-Cola								
51	Sprite								
52									
53									
54									
55									
56									
57									

Gráfico em Barras das Compras de Refrigerantes

Freqüência: 0, 5, 10, 15, 20

Coca-Cola | Coca-Cola Light | Dr. Pepper | Pepsi-Cola | Sprite

Refrigerante

Etapa 3. Quando o Assistente de Gráfico – Etapa 1 de 4 – Tipo de Gráfico aparecer:
Escolha **Coluna** na relação **Tipo de gráfico**
Escolha **Colunas Agrupadas** na janela **Subtipo de gráfico**
Dê um clique em **Avançar**

Etapa 4. Quando a caixa de diálogo Assistente de Gráfico – Etapa 2 de 4 – Dados de Origem aparecer:
Dê um clique em **Avançar**

Etapa 5. Quando a caixa de diálogo Assistente de Gráfico – Etapa 3 de 4 – Opções de Gráfico aparecer:
Selecione a guia **Títulos**
Digite Gráfico de Barras ou Compras de Refrigerantes na caixa **Título do Gráfico**
Digite Refrigerante na caixa **Eixo das Categorias (X)**
Digite Freqüência na caixa **Eixo dos Valores (Y)**
Selecione a guia **Legenda** e depois
Remova a marca de verificação da caixa **Mostrar legenda**
Dê um clique em **Avançar**

Etapa 6. Quando a caixa de diálogo Assistente de Gráfico – Etapa 4 de 4 – Localização do Gráfico aparecer:
Especifique uma localização para o novo gráfico (usamos a planilha atual ao selecionarmos **Como objeto em**)
Dê um clique em **Concluir**

O gráfico (diagrama) em barras resultante é mostrado na Figura 2.12.*

O Assistente de Gráfico do Excel pode produzir um gráfico em setores ("pizza") dos dados de compras de refrigerantes de maneira similar. Para desenvolver um gráfico em setores, escolha Pizza na relação Tipo de Gráfico da Etapa 3.

Conforme afirmamos no destaque "Notas e Comentários" no final da Seção 2.2, um histograma é fundamentalmente o mesmo que um gráfico em barras, sem nenhuma separação entre as barras. A Figura 2.13 mostra os dados de tempo para a conclusão das auditorias, com uma distribuição de freqüência em segundo plano e um gráfico em barras desenvolvido com o Assistente de Gráfico (usando as mesmas etapas que acabamos de descrever) em primeiro plano. Uma vez que as barras adjacentes de um histograma devem tocar-se, precisamos editar o gráfico de colunas (o gráfico em barras) a fim de eliminar o intervalo entre cada uma das barras. As etapas a seguir levam a efeito esse processo.

Etapa 1. Dê um clique com o botão direito do mouse em qualquer barra do gráfico de colunas para produzir uma lista de opções

Etapa 2. Escolha **Formatar Série de Dados**

Etapa 3 Quando a caixa de diálogo Formatar Série de Dados aparecer:
Selecione a guia **Opções**
Digite 0 na caixa **Largura do intervalo**
Dê um clique em **OK**

Diagrama de Dispersão Usamos os dados da loja de equipamentos de som da Tabela 2.13 para demonstrar o uso do Assistente de Gráfico do Excel para construir um diagrama de dispersão. Consulte a Figura 2.14 à medida que descrevermos as tarefas envolvidas. A planilha dos dados encontra-se em segundo plano e o diagrama de dispersão produzido pelo Assistente de Gráfico aparece em primeiro plano. As etapas a seguir produzirão o diagrama:

ARQUIVO DA INTERNET
Stereo

Etapa 1. Selecione as células B1:C11

Etapa 2. Selecione o botão **Assistente de Gráfico** na barra de ferramentas Padrão (ou selecione o menu **Inserir** e escolha a opção **Gráfico**)

Etapa 3. Quando a caixa de diálogo Assistente de Gráfico – Etapa 1 de 4 – Tipo de Gráfico aparecer:
Escolha **Dispersão (XY)** na janela **Tipo de gráfico:**
Dê um clique em **Avançar**

* Redimensionar um gráfico do Excel não é difícil. Primeiramente, selecione o gráfico. Pequenos quadrados, chamados alças de redimensionamento, surgirão nas bordas do gráfico. Dê um clique nas alças de redimensionamento e arraste-as para redimensionar a figura de acordo com sua preferência.

Etapa 4. Quando a caixa de diálogo Assistente de Gráfico – Etapa 2 de 4 – Dados de Origem aparecer:
Dê um clique em **Avançar**

Figura 2.13 Histograma dos dados de tempo para conclusão das auditorias construído com o Excel

	A	B	C	D	E	F	G	H	I	J	K
1	Tempo Audit.		Tempo Audit.	Freqüência							
2	12		10-14	4							
3	15		15-19	8							
4	20		20-24	5							
5	22		25-29	2							
6	14		30-34	1							
7	14										
8	15										
9	27										
10	21										
11	18										
12	19										
13	18										
14	22										
15	33										
16	16										
17	18										
18	17										
19	23										
20	28										
21	13										
22											
23											

Histograma dos Dados para a Conclusão das Auditorias
Freqüência
10
8
6
4
2
0
10-14
15-19
20-24
25-29
30-34
Tempo (para a conclusão) das auditorias

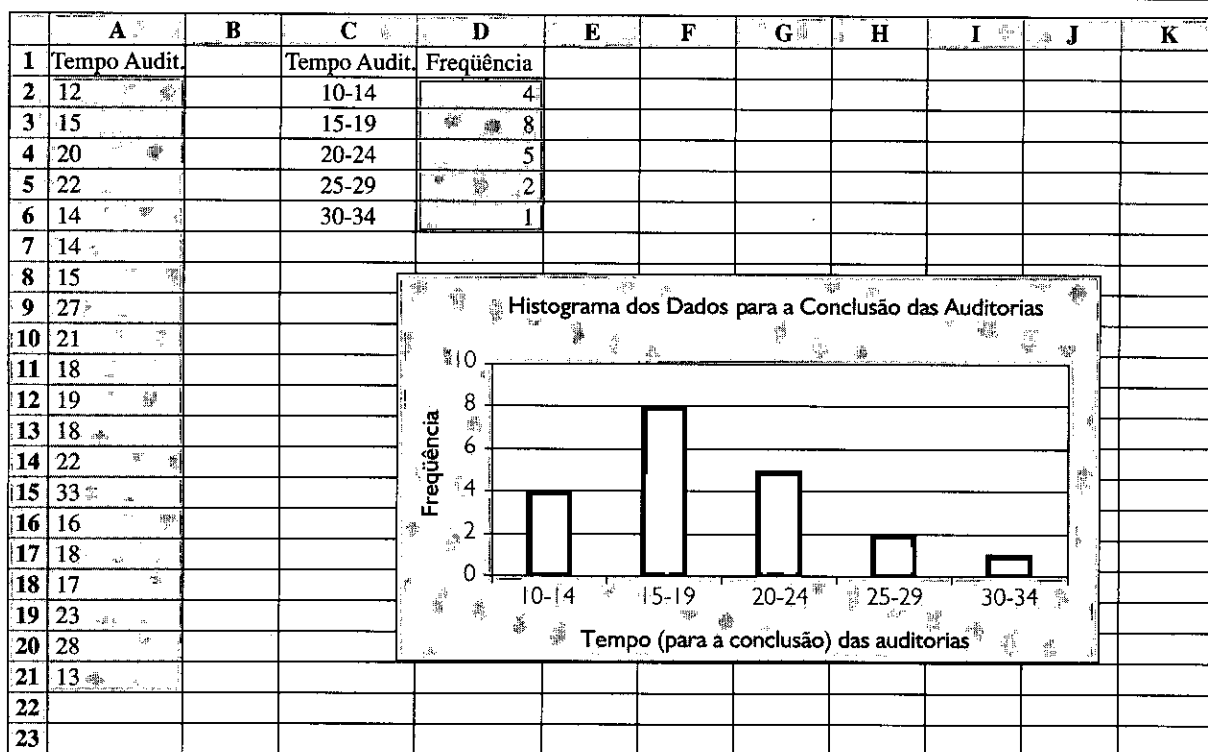



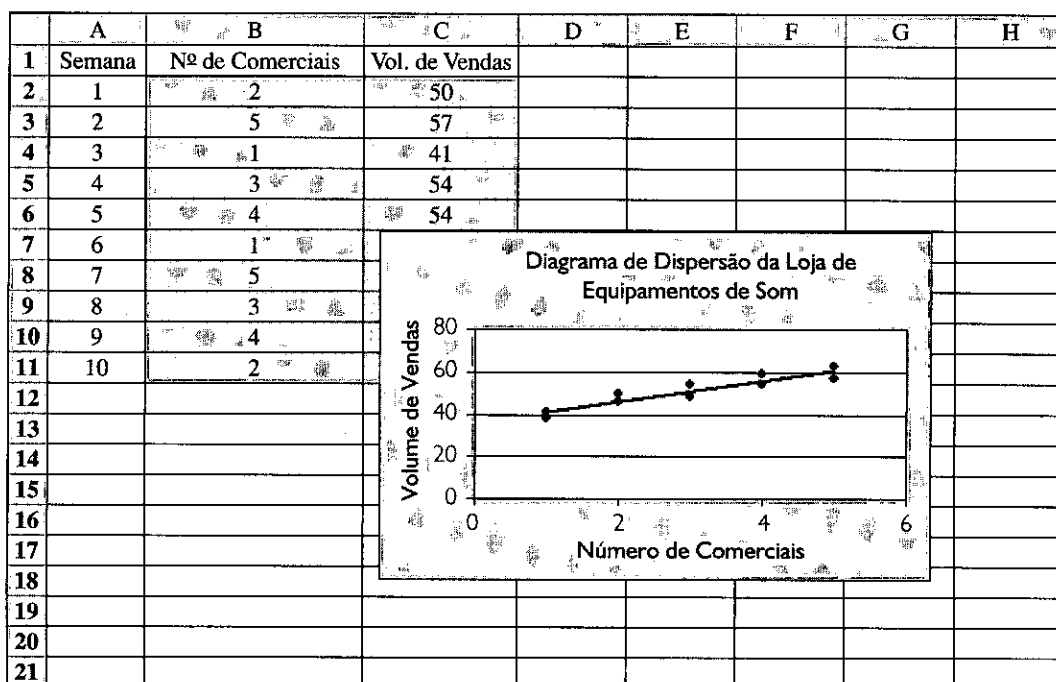
Figura 2.14 Diagrama de dispersão da loja de equipamentos de som criado com o assistente de gráfico do Excel

	A	B	C	D	E	F	G	H
1	Semana	Nº de Comerciais	Vol. de Vendas					
2	1	2	50					
3	2	5	57					
4	3	1	41					
5	4	3	54					
6	5	4	54					
7	6	1						
8	7	5						
9	8	3						
10	9	4						
11	10	2						
12								
13								
14								
15								
16								
17								
18								
19								
20								
21								

Diagrama de Dispersão da Loja de Equipamentos de Som
Volume de Vendas
80
60
40
20
0
0
2
4
6
Número de Comerciais

Etapa 5. Quando a caixa de diálogo Assistente de Gráfico – Etapa 3 de 4 – Opções de Gráfico aparecer:
Selecione a guia **Títulos**
Digite Diagrama de Dispersão da Loja de Equipamentos de Som na caixa **Título do gráfico**
Digite Comerciais na caixa **Eixo dos valores (X)**
Digite Volume de Vendas na caixa **Eixo dos Valores (Y)**
Selecione a guia **Legenda**
Remova a marca de verificação da caixa **Mostrar legenda**
Dê um clique em **Avançar**

Etapa 6. Quando a caixa de diálogo Assistente de Gráfico – Etapa 4 de 4 – Localização do Gráfico aparecer:
Especifique uma localização para o novo gráfico (Usamos a planilha atual ao selecionarmos **Como objeto em**)
Dê um clique em **Concluir**

Agora é fácil acrescentar uma linha de tendência ao diagrama de dispersão.

Etapa 1. Posicione o ponteiro do mouse sobre qualquer ponto de dados no diagrama de dispersão e dê um clique com o botão direito para exibir uma lista de opções

Etapa 2. Escolha **Adicionar Linha de Tendência**

Etapa 3. Quando a caixa de diálogo Adicionar Linha de Tendência aparecer:
Selecione a guia **Tipo**
Escolha **Linear** na janela **Tipo de Tendência/Regressão**
Dê um clique em **OK**

Relatório de Tabela Dinâmica

O Relatório de Tabela Dinâmica do Excel oferece uma ferramenta valiosa para gerenciar conjuntos de dados que envolvem mais de uma variável. Ilustraremos sua utilização mostrando como desenvolver uma tabulação cruzada.

Tabulação Cruzada Ilustramos a construção de uma tabulação cruzada usando os dados dos restaurantes apresentados na Figura 2.15. Os rótulos são introduzidos na linha 1 e os dados correspondentes a cada um dos 300 restaurantes são inseridos nas células A2:C301.

Etapa 1. Selecione o menu **Dados**

Etapa 2. Escolha **Relatório de Tabela e Gráfico Dinâmicos**

Etapa 3. Quando a caixa de diálogo Assistente de Tabela Dinâmica e Gráfico Dinâmico – Etapa 1 de 3 – aparecer:
Escolha **Banco de dados ou lista do Microsoft Office Excel**
Escolha **Tabela Dinâmica**
Dê um clique em **Avançar**

Etapa 4. Quando a caixa de diálogo Assistente de Tabela Dinâmica e Gráfico Dinâmico – Etapa 2 de 3 – aparecer:
Digite A1:C301 na caixa **Intervalo**
Dê um clique em **Avançar**

Etapa 5. Quando a caixa de diálogo Assistente de Tabela Dinâmica e Gráfico Dinâmico – Etapa 3 de 3 – aparecer:
Selecione **Na Nova Planilha**
Dê um clique em **Layout**

Etapa 6. Quando a caixa de diálogo Assistente de Tabela Dinâmica e Gráfico Dinâmico – Diagrama aparecer (veja a Figura 2.16):
Arraste o botão do campo **Avaliação da Qualidade** para a seção **LINHA** do diagrama
Arraste o botão do campo **Preço da Refeição (US$)** para a seção **COLUNA** do diagrama
Arraste o botão do campo **Restaurante** para a seção **DADOS** do diagrama
Dê um clique duplo sobre o botão do campo **Soma do Restaurante** na seção DADOS

ARQUIVO DA INTERNET
Restaurant

Nota: As linhas 12 a 291 estão ocultas.

Quando a caixa de diálogo Campo da Tabela Dinâmica aparecer:
Escolha **Contar** sob **Sintetizar por**
Dê um clique em **OK** (A Figura 2.17 mostra o diagrama concluído)
Dê um clique em **OK**

Figura 2.15 Planilha do Excel contendo dados dos restaurantes

	A	B	C	D
1	Restaurante	Aval. Qualidade	Preço da Refeição ($)	
2	1	Bom	18	
3	2	Ótimo	22	
4	3	Bom	28	
5	4	Excelente	38	
6	5	Ótimo	33	
7	6	Bom	28	
8	7	Ótimo	19	
9	8	Ótimo	11	
10	9	Ótimo	23	
11	10	Bom	13	
292	291	Ótimo	23	
293	292	Ótimo	24	
294	293	Excelente	45	
295	294	Bom	14	
296	295	Bom	18	
297	296	Bom	17	
298	297	Bom	16	
299	298	Bom	15	
300	299	Ótimo	38	
301	300	Ótimo	31	
302				

Figura 2.16 Assistente de tabela dinâmica e gráfico dinâmico – Diagrama

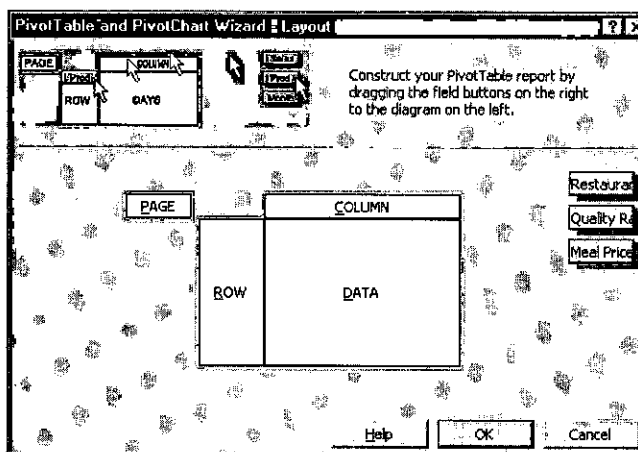
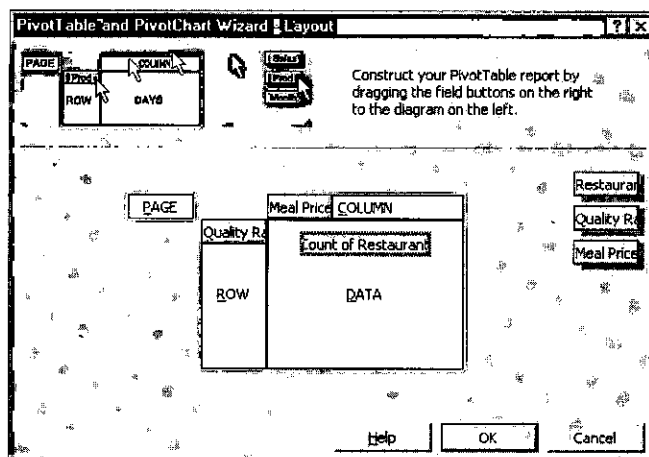



Figura 2.17 Diagrama concluído

Etapa 7. Quando a caixa de diálogo Assistente de Tabela Dinâmica e Gráfico Dinâmico – Etapa 3 de 3 – reaparecer:
Dê um clique em **Concluir**

Uma parte do *output*[15] gerado pelo Excel é mostrada na Figura 2.18. Note que o produto (*output*) que aparece nas colunas D a AK está oculto a fim de que os resultados possam ser visualizados em uma imagem razoavelmente grande. Os rótulos de linha (Excelente, Bom e Ótimo) e os totais de linha (66, 84, 150 e 300) que aparecem na Figura 2.18 são similares aos rótulos de linha e totais de linha expostos na Tabela 2.11. Mas eles estão em uma ordem diferente. Para colocá-los na ordem Bom, Ótimo e Excelente siga estas etapas.

Etapa 1. Dê um clique com o botão direito em **Excelente** na célula A5
Etapa 2. Escolha **Ordem**
Etapa 3. Selecione **Mover para o Fim**

Na Figura 2.18, uma coluna é designada para cada valor possível de preço das refeições. Por exemplo, a coluna B contém uma contagem dos restaurantes com preços de US$ 10 por refeição, a coluna C contém uma contagem dos restaurantes com preços de US$ 11 por refeição e assim por diante. Para visualizar o Relatório de Tabela Dinâmica de forma semelhante à mostrada na Tabela 2.11, devemos agrupar as colunas em quatro categorias de preços: US$ 10–19, US$ 20–29, US$ 30–39 e US$ 40–49. As etapas necessárias para agrupar as colunas correspondentes à planilha mostrada na Figura 2.18 são as seguintes:
Etapa 1. Dê um clique com o botão direito do mouse em Preço das Refeições (US$) na célula B3 da Tabela Dinâmica
Etapa 2. Escolha **Agrupar e Exibir Detalhe**
Escolha **Agrupar**
Etapa 3. Quando a caixa de diálogo **Agrupamento** aparecer:
Digite 10 na caixa **Iniciar em**
Digite 49 na caixa **Terminar em**
Digite 10 na caixa **Por**
Dê um clique em **OK**

O produto (*output*) revisado da Tabela Dinâmica é mostrado na Figura 2.19. É a Tabela Dinâmica final. Note que ela apresenta as mesmas informações que a tabulação cruzada exposta na Tabela 2.11.

[15] NT: *Output* – Dados de saída, resultado, produto (informática).

Figura 2.18 Resultado do relatório de tabela dinâmica inicial (as colunas D:AK estão ocultas)

	A	B	C	AL	AM	AN	AO
1							
2							
3	Contagem do Restaurante	Preço das Refeições (US$)					
4	Avaliação da Qualidade	10	11	47	48	Total Geral	
5	Excelente			2	2	66	
6	Bom	6	4			84	
7	Ótimo	1	4		1	150	
8	Total Geral	7	8	2	3	300	
9							
10							
11							
12							
13							
14							
15							
16							
17							
18							
19							
20							

PivotTable

Figura 2.19 Relatório de tabela dinâmica final dos dados dos restaurantes

	A	B	C	D	E	F	G
1							
2							
3	Contagem do Restaurante	Preço das Refeições (US$)					
4	Avaliação da Qualidade	10-19		20-29	30-39	40-49	Total Geral
5	Bom		42	40	2		84
6	Ótimo		34	64	46	6	150
7	Excelente		2	14	28	22	66
8	Total Geral		78	118	76	28	300
9							
10							
11							
12							
13							
14							
15							
16							
17							
18							
19							
20							

PivotTable

CAPÍTULO 3

Estatística Descritiva: Medidas Numéricas

ESTATÍSTICA NA PRÁTICA

SMALL FRY DESIGN*
Santa Ana, Califórnia

Fundada em 1997, a Small Fry Design é uma empresa de brinquedos e de acessórios que projeta e importa produtos para crianças. A linha de produtos da empresa inclui ursinhos de pelúcia, móbiles, brinquedos musicais, chocalhos e *security blankets*[1], caracterizando-se por projetos de brinquedos delicados de alta qualidade, com ênfase em cor, textura e som. Os produtos são projetados nos Estados Unidos e manufaturados na China.

A Small Fry Design utiliza representantes independentes para vender os produtos a varejistas que comercializam produtos infantis, lojas de roupas e acessórios para crianças, lojas de presentes, lojas de departamento de grande porte e grandes empresas que efetuam vendas por catálogo. Atualmente, os produtos da Small Fry Design são distribuídos em mais de mil canais de varejo em todo o território dos Estados Unidos.

O gerenciamento do fluxo de caixa é uma das atividades mais críticas nas operações diárias dessa empresa. Assegurar a suficiente entrada de caixa para satisfazer tanto as obrigações de débito atuais como as vindouras pode significar a diferença entre o sucesso e o fracasso do negócio. Um fator crucial no gerenciamento do

* Os autores agradecem a John A. McCarthy, presidente da Small Fry Design, por fornecer esta "Estatística na Prática".

[1] NT: *Security blanket*: um pequeno cobertor ou outro tecido macio ao qual as crianças se apegam ou no qual se envolvem devido à sensação de conforto e segurança que proporciona; qualquer coisa que dá a uma pessoa a sensação de segurança ou alívio da ansiedade.

fluxo de caixa é a análise e o controle das contas a receber. Ao calcular o período médio e o valor em dólares das faturas pendentes, os gerentes podem prever a disponibilidade de caixa e monitorar as alterações na posição das contas a receber. A empresa estabeleceu as seguintes metas: o tempo médio das faturas em haver não deve ultrapassar 45 dias e o valor em dólares das faturas com mais de 60 dias não deve exceder a 5% do valor em dólares de todas as contas a receber.

Em um sumário publicado recentemente a respeito da posição das contas a receber, foram apresentadas as seguintes estatísticas descritivas referentes ao tempo necessário para o recebimento das faturas:

Média	40 dias
Mediana	35 dias
Moda	31 dias

A interpretação dessas estatísticas mostra que a média, ou período médio, de uma fatura é de 40 dias. A mediana revela que metade das faturas permanece em haver durante 35 dias ou mais. A moda de 31 dias, que é o período mais freqüente das faturas, indica que 31 dias é a extensão de tempo mais comum que uma fatura permanece em haver. O sumário estatístico mostrou também que somente 3% do valor em dólares de todas as contas permaneceu acima de 60 dias. Tendo como base a informação estatística, a gerência convenceu-se de que as contas a receber e a entrada de caixa estavam sob controle.

Neste capítulo, você aprenderá a calcular e interpretar algumas das medidas estatísticas usadas pela Small Fry Design. Além da média, mediana e moda, você aprenderá outras estatísticas descritivas, por exemplo, amplitude, desvio padrão, percentis e correlação. Essas medidas numéricas vão ajudá-lo na compreensão e interpretação dos dados.

No Capítulo 2, discutimos os métodos tabulares e os métodos gráficos para sintetizar dados. Neste capítulo, apresentamos diversos métodos numéricos que constituem alternativas adicionais para sintetizar dados.

Iniciamos com o desenvolvimento de medidas numéricas resumidas de conjuntos de dados que consistem em uma única variável. Quando um conjunto de dados contém mais de uma variável, as mesmas medidas numéricas podem ser computadas separadamente para cada variável. Entretanto, no caso de duas variáveis, também desenvolveremos medidas da relação existente entre as variáveis.

Medidas numéricas de posição, dispersão, forma e associação serão apresentadas. Se as medidas calculadas referem-se aos dados de uma amostra, elas são chamadas **estatísticas da amostra**. Se as medidas calculadas referem-se a dados de uma população, elas são denominadas **parâmetros populacionais**. Em inferência estatística, uma estatística amostral refere-se a um **estimador por pontos** do parâmetro populacional correspondente. No Capítulo 7, discutiremos mais detalhadamente o processo de estimativa por pontos.

Nos dois apêndices deste capítulo, mostraremos como o Minitab e o Excel podem ser usados para calcularmos muitas das medidas numéricas aqui descritas.

3.1 MEDIDAS DE POSIÇÃO

Média

Talvez a medida de posição mais importante seja a **média**, ou valor médio, de uma variável. A média constitui uma medida da posição central dos dados. Se os dados se referem a uma amostra, a média é indicada por $\bar{x}$; se os dados correspondem a uma população, a média é indicada pela letra grega μ.

Nas fórmulas estatísticas é habitual exprimir-se o valor da variável x da primeira observação por x_1, o valor da variável x da segunda observação por x_2 e assim por diante. Em geral, o valor da variável x da i-ésima observação é indicado por x_i. Para uma amostra com n observações, a fórmula da média da amostra é a seguinte:

A média da amostra $\bar{x}$ é uma estatística da amostra.

MÉDIA DA AMOSTRA

$$\bar{x} = \frac{\Sigma x_i}{n} \tag{3.1}$$

Na fórmula anterior, o numerador é a soma dos valores das n observações. Ou seja,

$$\Sigma x_i = x_1 + x_2 + \cdots + x_n$$

A letra grega Σ é o símbolo de somatório.

Para ilustrar o cálculo de uma média da amostra, vamos considerar os seguintes dados de tamanho de classe de uma amostra de cinco classes universitárias:

46 54 42 46 32

Usamos a notação x_1, x_2, x_3, x_4 e x_5 para representar o número de estudantes em cada uma das cinco classes.

$$x_1 = 46 \quad x_2 = 54 \quad x_3 = 42 \quad x_4 = 46 \quad x_5 = 32$$

Portanto, para calcular a média da amostra, podemos escrever:

$$\bar{x} = \frac{\Sigma x_i}{n} = \frac{x_1 + x_2 + x_3 + x_4 + x_5}{5} = \frac{46 + 54 + 42 + 46 + 32}{5} = 44$$

O tamanho médio das classes da amostra é de 44 estudantes.

Outra ilustração do cálculo da média de uma amostra é dada na seguinte situação: suponha que o departamento de colocação profissional de uma universidade tenha enviado um questionário a uma amostra de diplomados da escola de administração, solicitando-lhes informações sobre salários mensais iniciais. A Tabela 3.1 mostra os dados coletados.

Tabela 3.1 Salários mensais iniciais de uma amostra de 12 graduados da escola de administração

Graduado	Salário Mensal Inicial (US$)	Graduado	Salário Mensal Inicial (US$)
1	2.850	7	2.890
2	2.950	8	3.130
3	3.050	9	2.940
4	2.880	10	3.325
5	2.755	11	2.920
6	2.710	12	2.880

ARQUIVO DA INTERNET

Salary

O salário mensal inicial médio da amostra de 12 graduados da escola de administração é calculado da seguinte maneira:

$$\bar{x} = \frac{\Sigma x_i}{n} = \frac{x_1 + x_2 + \cdots + x_{12}}{12}$$

$$= \frac{2.850 + 2.950 + \cdots + 2.880}{12}$$

$$= \frac{35.280}{12} = 2.940$$

A Equação 3.1 mostra como é calculada a média de uma amostra com n observações. A fórmula para calcular a média de uma população permanece a mesma, mas usamos uma notação diferente para indicar que estamos trabalhando com a população inteira. O número de observações em uma população é denotado por N, e o símbolo para a média de uma população é μ.

A média da amostra $\bar{x}$ é um estimador por pontos da média da população μ.

MÉDIA DA POPULAÇÃO

$$\mu = \frac{\Sigma x_i}{N} \tag{3.2}$$

Mediana

A **mediana** é outra medida da posição central de uma variável. A mediana é o valor intermediário quando os dados são organizados em ordem crescente (do menor valor para o maior valor). Quando se trata de um número ímpar de observações, a mediana é o valor intermediário. Um número par de observações não tem nenhum número intermediário em particular. Nesse caso, seguimos a convenção de definir a mediana como a média dos valores correspondentes às duas observações intermediárias. Por conveniência, a definição de mediana é reformulada e enunciada da seguinte maneira:

MEDIANA
Organize os dados em ordem crescente (do menor valor para o maior valor).
(a) Para um número ímpar de observações, a mediana é o valor intermediário.
(b) Para um número par de observações, a mediana é a média dos dois valores intermediários.

Vamos aplicar essa definição para calcular a mediana do tamanho de classe da amostra de cinco classes universitárias. Organizando os dados em ordem crescente, obtemos a seguinte lista:

32 42 46 46 54

Uma vez que o número de observações $n = 5$ é ímpar, a mediana é o valor intermediário. Desse modo, a mediana do tamanho das classes equivale a 46 estudantes. Embora esse conjunto de dados contenha duas observações com valores 46, cada observação é tratada separadamente quando organizamos os dados em ordem crescente.

Suponha também que calculemos a mediana do salário inicial dos 12 graduados da escola de administração da Tabela 3.1. Primeiramente, organizamos os dados em ordem crescente:

2.710 2.755 2.850 2.880 2.880 2.890 2.920 2.940 2.950 3.050 3.130 3.325

Os dois valores intermediários

A mediana é a medida de posição mais freqüentemente usada para dados de renda anual e valor patrimonial porque algumas rendas ou valores patrimoniais extremamente elevados podem inflacionar a média. Nesses casos, a mediana é a medida preferível da posição central.

Já que $n = 12$ é par, identificamos os dois valores intermediários: 2.890 e 2.920. A mediana é a média desses valores.

$$\text{Mediana} = \frac{2.890 + 2.920}{2} = 2.905$$

Não obstante a média ser a medida de posição central mais comumente usada, em algumas situações é preferível usar a mediana. A média é influenciada por valores de dados extremamente pequenos ou grandes. Por exemplo, suponha que um dos graduados (veja a Tabela 3.1) tenha um salário inicial de US$ 10.000 por mês (talvez a família dessa pessoa seja a dona da empresa). Se mudarmos o salário mensal inicial mais elevado da Tabela 3.1 de US$ 3.325 para US$ 10.000 e recalcularmos a média, a média da amostra passará de US$ 2.940 para US$ 3.496. A mediana de US$ 2.905, entretanto, não se alterará, porque US$ 2.890 e US$ 2.920 ainda são os valores intermediários. Ao incluirmos o salário inicial extremamente elevado, a mediana nos fornece uma medida mais acurada da posição central do que a média. Podemos generalizar e afirmar que, quando um conjunto de dados contém valores extremos, freqüentemente a mediana é a medida de posição central preferível.

Moda

Uma terceira medida da posição é a **moda**. A moda é definida da seguinte maneira:

MODA
Moda é o valor que ocorre com maior freqüência.

Para ilustrar a identificação da moda, considere a amostra de cinco tamanhos de classe. O único valor que ocorre mais de uma vez é 46. Uma vez que esse valor tem a maior freqüência, pois ocorre duas vezes, ele é a moda. Como outra ilustração, considere a amostra de salários iniciais dos graduados da escola de administração. O único salário mensal inicial que ocorre mais de uma vez é US$ 2.880. Uma vez que esse valor tem a maior freqüência, ele é a moda.

Pode haver situações em que a maior freqüência ocorre em dois ou mais valores diferentes. Nesses casos, existe mais de uma moda. Se os dados têm exatamente duas modas, dizemos que os dados são *bimodais*. Se os dados possuem mais de duas modas, os denominamos *multimodais*. Nos casos multimodais, a moda quase nunca é considerada, porque relacionar três ou mais modas não seria especialmente útil para descrever a posição dos dados.

A moda é uma medida importante da posição de dados qualitativos. Por exemplo, o conjunto dos dados qualitativos da Tabela 2.2 resultou na seguinte distribuição de freqüência das compras de refrigerantes:

Refrigerante	Freqüência
Coca-Cola	19
Coca-Cola Light	8
Dr. Pepper	5
Pepsi-Cola	13
Sprite	5
Total	50

A moda, ou o refrigerante mais comprado, é a Coca-Cola. Para esse tipo de dados, evidentemente, não tem sentido falarmos em moda ou mediana. A moda fornece a informação que nos interessa: o refrigerante comprado com maior freqüência.

Percentis

Um **percentil** fornece a informação sobre como os dados se distribuem ao longo do intervalo entre o menor e o maior valor. Para dados que não têm muitos valores repetidos, o p-ésimo percentil divide os dados em duas partes. Aproximadamente p por cento das observações apresentam valores menores que o p-ésimo percentil; aproximadamente $(100 - p)$ por cento das observações possuem valores maiores que o p-ésimo percentil. O p-ésimo percentil é formalmente definido da seguinte maneira:

PERCENTIL

O p-ésimo percentil é um valor tal que *pelo menos p* por cento das observações são menores ou iguais a esse valor e *pelo menos* $(100 - p)$ por cento das observações são maiores ou iguais a esse valor.

Colégios e universidades geralmente registram notas de exames de admissão em termos de percentis. Por exemplo, suponha que um candidato obtenha a nota bruta de 54 pontos na parte oral de um exame de admissão. O desempenho desse estudante em relação a outros estudantes que fizeram o mesmo exame pode não ser claro imediatamente. Entretanto, se a nota bruta de 54 pontos corresponde ao 70º percentil, sabemos que aproximadamente 70% dos estudantes tiveram pontuações menores que esse indivíduo e que aproximadamente 30% dos estudantes tiveram notas mais altas do que ele.

O procedimento a seguir pode ser usado para calcular o p-ésimo percentil:

PARA CALCULAR O p-ÉSIMO PERCENTIL

Etapa 1: Organize os dados em ordem crescente (do menor valor para o maior valor).

Etapa 2: Calcule um índice i

$$i = \left(\frac{p}{100}\right)n$$

em que p é o percentil procurado e n, o número de observações.

Etapa 3: (a) *Se* i *não for um número inteiro*, arredonde-o para cima. O número inteiro seguinte *maior* que i denota a posição do p-ésimo percentil.

(b) *Se* i *for um número inteiro*, o p-ésimo percentil será a média dos valores nas posições i e $i + 1$.

Depois dessas etapas, torna-se mais fácil calcular o percentil.

Como ilustração desse procedimento, vamos determinar o 85º percentil dos dados de salários iniciais da Tabela 3.1.

Etapa 1: Organize os dados em ordem crescente.

2.710 2.755 7.850 2.880 2.880 2.890 2.920 2.940 2.950 3.050 3.130 3.325

Etapa 2:

$$i = \left(\frac{p}{100}\right)n = \left(\frac{85}{100}\right)12 = 10,2$$

Etapa 3: Uma vez que i não é um número inteiro, *arredonde-o para cima.* A posição do 85º percentil é o número inteiro seguinte maior que 10,2, a 11ª posição.

Retornando aos dados, vemos que o 85º percentil é o valor de dados que está na 11ª posição, ou seja, 3.130.

Como outra ilustração desse procedimento, consideremos o cálculo do 50º percentil dos dados de salários iniciais. Aplicando a etapa 2, obtemos:

$$i = \left(\frac{50}{100}\right)12 = 6$$

Uma vez que i é um número inteiro, a etapa 3(b) afirma que o 50º percentil é a média do sexto e sétimo valores de dados; dessa forma, o 50º percentil é (2.890 + 2.920)/2 = 2.905. Observe que o *50º percentil é também a mediana.*

Quartis são apenas percentis específicos; desse modo, as etapas para calcular percentis podem ser aplicadas diretamente no cálculo dos quartis.

Quartis

Muitas vezes é desejável dividir os dados em quatro partes, tendo cada parte aproximadamente um quarto, ou 25% das observações. A Figura 3.1 mostra uma distribuição de dados dividida em quatro partes. Os pontos da divisão denominam-se **quartis** e são definidos como:

Q_1 = o primeiro quartil, ou 25º percentil
Q_2 = o segundo quartil, ou 50º percentil (também, a mediana)
Q_3 = o terceiro quartil, ou 75º percentil

Os dados dos salários iniciais são novamente organizados em ordem crescente. Já identificamos Q_2, o segundo quartil (mediana), como 2.905.

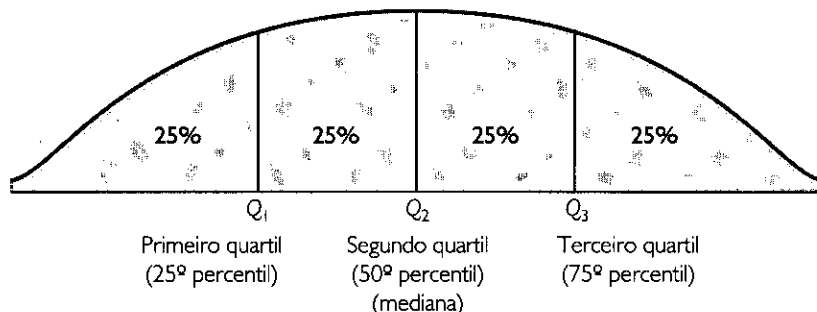
2.710 2.755 7.850 2.880 2.880 2.890 2.920 2.940 2.950 3.050 3.130 3.325

O cálculo dos quartis Q_1 e Q_2 requer o uso da regra aplicada para se encontrar o 25º e o 75º percentis. Os cálculos são os seguintes:

Para Q_1,

$$i = \left(\frac{p}{100}\right)n = \left(\frac{25}{100}\right)12 = 3$$

Figura 3.1 Posição dos quartis

25% 25% 25% 25%
Q_1 Primeiro quartil (25º percentil)
Q_2 Segundo quartil (50º percentil) (mediana)
Q_3 Terceiro quartil (75º percentil)

Visto que i é um número inteiro, a etapa 3(b) indica que o primeiro quartil, ou o 25º percentil, é a média do terceiro e quarto valores de dados; desse modo, Q_1 = (2.850 + 2.880)/2 = 2.865.

Para Q_3,

$$i = \left(\frac{p}{100}\right)n = \left(\frac{75}{100}\right)12 = 9$$

Novamente, desde que *i* é um número inteiro, a etapa 3(b) indica que o terceiro quartil, ou o 75º percentil, é a média do nono e décimo valores de dados; assim, $Q_3 = (2.950 + 3.050)/2 = 3.000$.

Os quartis dividem os dados de salários iniciais em quatro partes, e cada parte contém 25% das observações.

2.710 2.755 2.850	2.880 2.880 2.890	2.920 2.940 2.950	3.050 3. 130 3.325

$Q_1 = 2865$ $Q_2 = 2905$ (Mediana) $Q_3 = 3000$

Definimos os quartis como o 25º, o 50º e o 75º percentis. Assim, calculamos os quartis da mesma maneira que calculamos os percentis. Entretanto, às vezes são usadas outras convenções para se calcular os quartis, e os valores reais atribuídos aos quartis podem variar ligeiramente, dependendo da convenção utilizada. Contudo, o objetivo de todos os procedimentos para se calcular quartis é dividir os dados em quatro partes iguais.

NOTAS E COMENTÁRIOS

Quando um conjunto de dados contém valores extremos, é melhor usar a mediana, em vez da média, como medida da posição central. Outra medida, às vezes usada quando se tem valores extremos, é a *média ajustada*. Ela é obtida excluindo-se uma porcentagem dos valores menores e maiores de um conjunto de dados e calculando-se então a média dos valores restantes. Por exemplo, a média ajustada de 5% é obtida eliminando-se os 5% dos valores de dados menores e os 5% dos valores de dados maiores e calculando-se depois a média dos valores restantes. Ao usarmos a amostra com $n = 12$ salários iniciais, teremos 0,05(12) = 0,6. O arredondamento desse valor para 1 indica que a média ajustada de 5% significaria eliminar o menor valor de dados e o maior valor de dados. A média ajustada de 5%, usando-se as dez observações restantes, é 2.924,50.

Exercícios

Métodos

1. Considere uma amostra com os valores de dados 10, 20, 12, 17 e 16. Calcule a média e a mediana.
2. Considere uma amostra com os valores de dados 10, 20, 21, 17, 16 e 12. Calcule a média e a mediana.
3. Considere uma amostra com os valores de dados 27, 25, 20, 15, 30, 34, 28 e 25. Calcule o 20º, o 25º, o 65º e o 75º percentis.

AUTOTESTE

4. Considere uma amostra com os valores de dados 53, 55, 70, 58, 64, 57, 53, 69, 57, 68 e 53. Calcule a média, a mediana e a moda.

Aplicações

5. A publicação Dow Jones Travel Index divulgou o valor que as pessoas que viajam a negócios pagam por pernoite em quartos de hotel nas principais cidades dos Estados Unidos (*The Wall Street Journal*, 16 de janeiro de 2004). A média dos preços de quartos de hotel de 20 cidades são as seguintes:

Atlanta	$163	Minneapolis	$125
Boston	177	Nova Orleans	167
Chicago	166	Nova York	245
Cleveland	126	Orlando	146
Dallas	123	Phoenix	139
Denver	120	Pittsburgh	134
Detroit	144	São Francisco	167
Houston	173	Seattle	162
Los Angeles	160	St. Louis	145
Miami	192	Washington, D.C.	207

ARQUIVO DA INTERNET

Hotel

a. Qual é a média dos preços de quartos de hotel?
b. Qual é a mediana dos preços de quartos de hotel?
c. Qual é a moda?
d. Qual é o primeiro quartil?
e. Qual é o terceiro quartil?

6. A J. D. Powers and Associates fez uma pesquisa de usuários de telefones celulares a fim de saber quantos minutos eles usavam telefones celulares por mês (*Associated Press*, junho de 2002). Os minutos por mês relativos a uma amostra de 15 usuários de telefones celulares são mostrados a seguir:

615	135	395
430	830	1180
690	250	420
265	245	210
180	380	105

a. Qual é a média de minutos de uso por mês?
b. Qual é a mediana de minutos de uso por mês?
c. Qual é o 85º percentil?
d. A J. D. Powers and Associates divulgou que a média dos planos de assinatura de telefones sem fio permite até 750 minutos de uso por mês. O que esses dados sugerem a respeito da utilização que os assinantes de telefones celulares fazem de seus planos de assinatura mensal?

7. A American Association of Individual Investors realizou uma pesquisa anual de *discount brokers*[2] (*AAII Journal*, janeiro de 2003). As comissões cobradas pelas 24 *discount brokers* para dois tipos de transações, a comercialização de 100 ações a US$ 50 por ação auxiliada por corretores e a comercialização on-line de 500 ações a US$ 50 por ação, são mostradas a Tabela 3.2.

ARQUIVO DA INTERNET
Broker

Tabela 3.2 Comissões cobradas pelas *discount brokers*

Corretora	Comercialização de 100 Ações a US$ 50 por Ação Auxiliada por Corretores	Comercialização On-Line de 500 Ações a US$ 50 por Ação	Corretora	Comercialização de 100 Ações a US$ 50 por Ação Auxiliada por Corretores	Comercialização On-Line de 500 Ações a US$ 50 por Ação
Accutrade	30,00	29,95	Merrill Lynch Direct	50,00	29,95
Ameritrade	24,99	10,99	Muriel Siebert	45,00	14,95
Banc of America	54,00	24,95	NetVest	24,00	14,00
Brown & Co,	17,00	5,00	Recom Securities	35,00	12,95
Charles Schwab	55,00	29,95	Scottrade	17,00	7,00
CyberTrader	12,95	9,95	Sloan Securities	39,95	19,95
E*TRADE Securities	49,95	14,95	Strong Investments	55,00	24,95
First Discount	35,00	19,75	TD Waterhouse	45,00	17,95
Freedom Investments	25,00	15,00	T. Rowe Price	50,00	19,95
Harrisdirect	40,00	20,00	Vanguard	48,00	20,00
Investors National	39,00	62,50	Wall Street Discount	29,95	19,95
MB Trading	9,95	10,55	York Securities	40,00	36,00

Fonte: AAII Journal, janeiro de 2003.

a. Calcule a média, a mediana e a moda da comissão cobrada na comercialização de 100 ações a US$ 50 por ação auxiliada por corretores.
b. Calcule a média, a mediana e a moda da comissão cobrada na comercialização on-line de 500 ações a US$ 50 por ação.
c. O que custa mais: a comercialização de 100 ações a US$ 50 por ação auxiliada por corretores ou a comercialização on-line de 500 ações a US$ 50 por ação?
d. O custo de uma transação se relaciona com o valor da transação?

[2] NT: *Discount broker* – As corretoras chamadas *discount brokers*, ou de descontos, oferecem serviços de operação financeira (compra e venda de futuros e opções da bolsa de valores) com foco na agilidade e na prática de preços. Elas apenas executam as ordens dos clientes, sem análise de papéis (economia).

8. Milhões de norte-americanos se levantam de manhã e realizam seu trabalho em escritórios residenciais, comunicando-se com a empresa por meios eletrônicos. Apresentamos, a seguir, uma amostra de dados de faixa etária de indivíduos que trabalham em casa:

AUTOTESTE

18	54	20	46	25	48	53	27	26	37
40	36	42	25	27	33	28	40	45	25

a. Calcule a média e a moda.
b. A mediana da idade da população de todos os adultos é 35,5 anos (*The World Almanac*, 2004). Use a mediana de idade dos dados anteriores para comentar se as pessoas que trabalham em casa tendem a ser mais jovens ou mais velhas que a população de todos os adultos.
c. Calcule o primeiro e o terceiro quartis.
d. Calcule e interprete o 32º percentil.

9. A Media Matrix coletou dados que mostram os *websites* mais populares quando se navega em casa e no trabalho (*Business 2.0*, janeiro de 2000). Os dados a seguir mostram o número de visitantes exclusivos (em milhares) dos 25 sites mais freqüentados quando se navega em casa:

ARQUIVO DA INTERNET
Websites

Website	Visitantes Exclusivos (em milhares)
about.com	5.538
altavista.com	7.391
amazon.com	7.986
angelfire.com	8.917
aol.com	23.863
bluemountainarts.com	6.786
ebay.com	8.296
excite.com	10.479
geocities.com	15.321
go.com	14.330
hotbot.com	5.760
hotmail.com	11.791
icq.com	5.052
looksmart.com	5.984
lycos.com	9.950
microsoft.com	15.593
msn.com	23.505
netscape.com	14.470
passport.com	11.299
real.com	6.785
snap.com	5.730
tripod.com	7.970
xoom.com	5.652
yahoo.com	26.796
zdnet.com	5.133

a. Calcule a média e a mediana.
b. Você acha que seria melhor usar a média ou a mediana como medida da posição central desses dados? Explique.
c. Calcule o primeiro e o terceiro quartis.
d. Calcule e interprete o 85º percentil.

10. Uma pesquisa realizada pela American Hospital Association descobriu que as salas de emergência da maioria dos hospitais operam em plena capacidade (Associated Press, 9 de abril de 2002). A pesquisa coletou dados sobre os tempos de espera para as salas de emergência dos hospitais nos quais elas funcionam em plena capacidade e dos hospitais nos quais as salas de emergência encontram-se em equilíbrio e raramente operam em sua plena capacidade. Os dados de amostra que apresentam os tempos de espera em minutos são os seguintes:

Tempos de Espera para as Salas de Emergência dos Hospitais nos quais Elas Funcionam em Plena Capacidade		Tempos de Espera para as Salas de Emergência dos Hospitais nos quais Elas Estão em Equilíbrio	
87	59	60	39
80	110	54	32
47	83	18	56
73	79	29	26
50	50	45	37
93	66	34	38
72	115		

a. Calcule a média e a mediana dos tempos de espera para as salas de emergência de hospitais nos quais elas funcionam em sua plena capacidade.
b. Calcule a média e a mediana dos tempos de espera para as salas de emergência dos hospitais nos quais elas estão em equilíbrio.
c. Quais observações você é capaz de fazer a respeito dos tempos de espera para as salas de emergência baseando-se nesses resultados? A American Hospital Association expressaria alguma preocupação com os resultados estatísticos aqui mostrados?

11. Em um teste automobilístico de quilometragem e consumo de gasolina, 13 automóveis foram testados na estrada, em um percurso de 482,80 quilômetros, em condições de dirigibilidade tanto na cidade como na rodovia. Os dados apresentados a seguir foram registrados para o desempenho obtido em termos de quilômetros por galão.[3]

Cidade	26,07	26,87	25,58	23,17	21,24	24,62	27,03	25,74	25,91	24,62	24,46	24,62	25,74
Rodovia	30,57	32,18	28,96	29,93	30,89	27,35	27,35	28,96	30,57	33,95	31,22	28,96	28,96

Use a média, a mediana e a moda para fazer uma afirmação sobre a diferença de desempenho quando se dirige na cidade e na rodovia.

12. Os dados apresentados a seguir mostram o preço, a capacidade de imagem e o tempo de duração da bateria (em minutos) de 20 câmeras digitais (*PC World*, janeiro de 2000):

ARQUIVO DA INTERNET
Cameras

Câmera	Preço (US$)	Capacidade de Imagem	Duração da Bateria (em minutos)
Agfa Ephoto CL30	349	36	25
Canon PowerShot A50	499	106	75
Canon PowerShot Pro70	999	96	118
Epson PhotoPC 800	699	120	99
Fujifilm DX-10	299	30	229
Fujifilm MX-2700	699	141	124
Fujifilm MX-2900 Zoom	899	141	88
HP PhotoSmart C200	299	80	68
Kodak DC215 Zoom	399	54	159
Kodak DC265 Zoom	899	180	186
Kodak DC280 Zoom	799	245	143
Minolta Dimage EX Zoom 1500	549	105	38
Nikon Coolpix 950	999	32	88
Olympus D-340R	299	122	161
Olympus D-450 Zoom	499	122	62
Ricoh RDC-500	699	99	56
Sony Cybershot DSC-F55	699	63	69
Sony Mavica MVC-FD73	599	40	186
Sony Mavica MVC-FD88	999	40	88
Toshiba PDR-M4	599	124	142

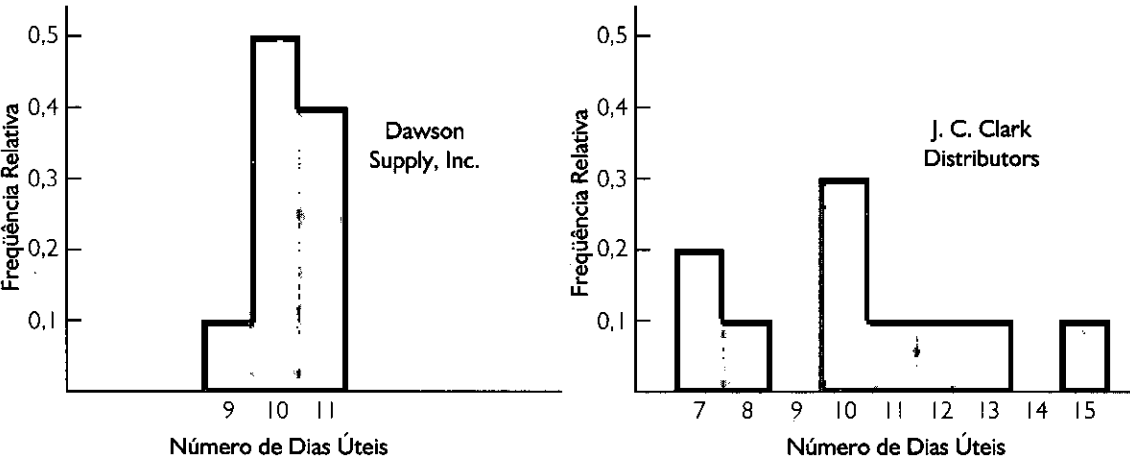
a. Calcule o preço médio.
b. Calcule a média de capacidade de imagem.
c. Calcule a média do tempo de duração da bateria.
d. Se você tivesse de escolher uma câmera dessa lista, qual delas escolheria? Explique.

[3] NT: Galão: 3,78 litros.

3.2 MEDIDAS DE VARIABILIDADE

Além das medidas de posição, freqüentemente é desejável levarmos em consideração as medidas de variabilidade, ou de dispersão. Por exemplo, suponha que você seja um agente de compras de uma grande empresa de manufatura e que regularmente faça pedidos de compra a dois fornecedores diferentes. Depois de vários meses de operação, você descobre que o número médio de dias necessários para aviarem os pedidos é de dez dias para ambos os fornecedores. Os histogramas que sintetizam o número de dias úteis necessários para que os fornecedores aviem os pedidos são mostrados na Figura 3.2. Não obstante o número médio de dias ser dez para ambos os fornecedores, os dois fornecedores demonstram o mesmo grau de confiabilidade em termos de efetuarem as entregas no prazo devido? Note a dispersão, ou a variabilidade, dos prazos de entrega, indicada pelos histogramas. Qual fornecedor você preferiria?

Para a maioria das empresas, receber matérias-primas e suprimentos no prazo programado é importante. Os prazos de entrega de sete ou oito dias mostrados para a J. C. Clark Distributors poderiam ser vistos favoravelmente; entretanto, algumas das entregas que se retardam de 13 a 15 dias poderiam ser desastrosas em termos de manter a mão-de-obra ocupada e a produção dentro do prazo determinado.

Figura 3.2 Dados históricos com o número de dias necessários para o aviamento dos pedidos de compra

Esse exemplo ilustra uma situação na qual a variabilidade nos prazos de entrega pode ter uma importância fundamental na escolha de um fornecedor. Para a maioria dos agentes de compra, a menor variabilidade apresentada pela Dawson Supply, Inc. tornaria esse fornecedor o preferível.

Voltamo-nos agora à discussão de algumas medidas de variabilidade comumente usadas.

Amplitude

A medida mais simples de variabilidade é a **amplitude**.

AMPLITUDE

$$\text{Amplitude} = \text{Maior valor} - \text{Menor valor}$$

Consultemos os dados sobre salários iniciais dos graduados da escola de administração apresentados na Tabela 3.1. O maior salário inicial é 3.325 e o menor, 2.710. A amplitude é 3.325 – 2.710 = 615.

Ainda que a amplitude seja a medida de variabilidade mais fácil de calcular, raramente ela é usada como a única medida. A razão para isso é que a amplitude se baseia somente em duas das observações e, desse modo, é altamente influenciada por valores extremos. Suponha que um dos graduados receba um salário inicial de US$ 10.000 por mês. Nesse caso, a amplitude seria 10.000 – 2.710 = 7.290 em vez de 615. Esse valor elevado para a amplitude não descreveria de maneira especial a variabilidade dos dados, porque 11 dos 12 salários iniciais estão estreitamente agrupados entre 2.710 e 3.130.

Amplitude Interquartil

Uma medida da variabilidade que supera a dependência de valores extremos é a **amplitude interquartil (AIQ)**. Essa medida da variabilidade é a diferença entre o terceiro quartil, Q_3, e o primeiro quartil, Q_1. Em outras palavras, a amplitude interquartil é o intervalo correspondente aos 50% dos dados intermediários.

AMPLITUDE INTERQUARTIL

$$AIQ = Q_3 - Q_1 \tag{3.3}$$

Em relação aos dados sobre salários mensais iniciais, os quartis são $Q_3 = 3.000$ e $Q_1 = 2.865$. Desse modo, a amplitude interquartil é 3.000 – 2.865 = 135.

Variância

Variância é a medida da variabilidade que utiliza todos os dados. A variância baseia-se na diferença entre o valor de cada observação (x_i) e a média. A diferença entre cada x_i e a média ($\bar{x}$ para uma amostra e μ para uma população) denomina-se *desvio em torno da média*. Para uma amostra, o desvio em torno da média é escrito como $(x_i - \bar{x})$; para uma população, ele é escrito como $(x_i - \mu)$. No cálculo da variância, os desvios em torno da média são *elevados ao quadrado*.

Se os dados se referirem a uma população, a média dos desvios elevados ao quadrado denomina-se *variância da população*. A variância da população é denotada pelo símbolo grego σ^2. Para uma população de N observações, com μ denotando a média da população, a definição da variância da população é a seguinte:

VARIÂNCIA DA POPULAÇÃO

$$\sigma^2 = \frac{\Sigma(x_i - \mu)^2}{N} \tag{3.4}$$

Na maioria das aplicações estatísticas, os dados que são analisados referem-se a uma amostra. Quando calculamos a variância de uma amostra, freqüentemente nos interessa usá-la para estimar a variância da população σ^2. Não obstante a explicação detalhada estar além do escopo deste livro, pode-se demonstrar que se a soma dos desvios em torno da média da amostra elevados ao quadrado for dividida por $n - 1$, e não por n, a resultante variância da amostra fornecerá uma estimativa sem tendenciosidade da variância da população. Por essa razão, a *variância da amostra*, denotada por s^2, é definida da seguinte maneira:

A variância da amostra s^2 é a estimativa da variância populacional σ^2.

VARIÂNCIA DA AMOSTRA

$$s^2 = \frac{\Sigma(x_i - \bar{x})^2}{n - 1} \tag{3.5}$$

Para ilustrar o cálculo da variância da amostra usaremos os dados dos tamanhos de classe da amostra de cinco classes universitárias apresentados na Seção 3.1. Um resumo dos dados, incluindo o cálculo dos desvios em torno da média e os desvios em torno da média elevados ao quadrado, é mostrado na Tabela 3.3. A soma dos desvios em torno da média elevados ao quadrado é $\Sigma(x_i - \bar{x})^2 = 256$. Portanto, com $n - 1 = 4$, a variância da amostra é

$$s^2 = \frac{\Sigma(x_i - x)^2}{n - 1} = \frac{256}{4}$$

Antes de prosseguir, vamos notar que as unidades associadas à variância da amostra muitas vezes causam confusão. Uma vez que os valores que são somados no cálculo da variância, $(x_i - \bar{x})^2$, estão elevados ao quadrado, as unidades associadas à variância da amostra também são *elevadas ao quadrado*. Por exemplo, a variância da amostra dos dados de tamanhos de classe é $s^2 = 64$ (estudantes)^2.

Tabela 3.3 Cálculo dos desvios e dos desvios em torno da média elevados ao quadrado dos dados de tamanhos de classe

Número de Estudantes na Classe (x_i)	Média da Amostra ($\bar{x}$)	Desvio em Torno da Média ($x_i - \bar{x}$)	Desvio em Torno da Média Elevado ao Quadrado ($x_i - \bar{x})^2$
46	44	2	4
54	44	10	100
42	44	22	4
46	44	2	4
32	44	212	144
		0	256
		$\Sigma(x_i - \bar{x})$	$\Sigma(x_i - x)^2$

As unidades elevadas ao quadrado associadas à variância tornam difícil obter uma compreensão e uma interpretação intuitivas do valor numérico da variância. Recomendamos que você considere a variância uma medida útil ao comparar a quantidade de variabilidade de duas ou mais variáveis. Em uma comparação de variáveis, aquela que tem a maior variância exibe mais variabilidade. Uma interpretação adicional do valor da variância talvez não seja necessária.

A variância é útil para comparar a variabilidade de duas ou mais variáveis.

Como outra ilustração do cálculo de uma variância da amostra, considere os salários iniciais relacionados na Tabela 3.1 para os 12 graduados da escola de administração. Na Seção 3.1 mostramos que a média dos salários iniciais da amostra era 2940. O cálculo da variância da amostra (s^2 = 27.440,91) é mostrado na Tabela 3.4.

Tabela 3.4 Cálculo da variância da amostra dos dados de salários iniciais

Salário Mensal (x_i)	Média da Amostra ($\bar{x}$)	Desvio em Torno da Média ($x_i - \bar{x}$)	Desvio em Torno da Média Elevado ao Quadrado ($x_i - \bar{x})^2$
2.850	2.940	−90	8.100
2.950	2.940	10	100
3.050	2.940	110	12.100
2.880	2.940	−60	3.600
2.755	2.940	−185	34.225
2.710	2.940	−230	52.900
2.890	2.940	−50	2.500
3.130	2.940	190	36.100
2.940	2.940	0	0
3.325	2.940	385	148.225
2.920	2.940	−20	400
2.880	2.940	−60	3.600
		0	301,850
		$\Sigma(x_i - \bar{x})$	$\Sigma(x_i - \bar{x})^2$

Usando a Equação 3.5,

$$s^2 = \frac{(x_i - x)^2}{n - 1} = \frac{30.850}{11} = 27.440,91$$

Nas Tabelas 3.3 e 3.4 mostramos tanto a soma dos desvios em torno da média como a soma dos desvios em torno da média elevados ao quadrado. Para qualquer conjunto de dados, a soma dos desvios em torno da média *sempre será igual a zero*. Note que nas Tabelas 3.3 e 3.4, $\Sigma(x_i - \bar{x}) = 0$. Os desvios positivos e os desvios negativos se cancelam mutuamente, fazendo que a soma dos desvios em torno da média seja igual a zero.

Desvio Padrão

O **desvio padrão** é definido como a raiz quadrada positiva da variância. Seguindo a notação que adotamos para uma variância da amostra e para uma variância da população, usamos s para denotar o desvio padrão da amostra e s para denotar o desvio padrão da população. O desvio padrão é derivado da variância da seguinte maneira:

O desvio padrão da amostra s é o estimador do desvio padrão da população σ.

DESVIO PADRÃO

$$\text{Desvio padrão da amostra} = s = \sqrt{s^2} \tag{3.6}$$

$$\text{Desvio padrão da população} = \sigma = \sqrt{\sigma^2} \tag{3.7}$$

O desvio padrão é mais fácil de interpretar que a variância porque ele é medido nas mesmas unidades que os dados.

Lembre-se de que a variância da amostra referente à amostra de tamanhos de classe de cinco classes universitárias é $s^2 = 64$. Desse modo, o desvio padrão da amostra é $s = \sqrt{64} = 8$. Em relação aos dados sobre salários iniciais, o desvio padrão da amostra é $s = \sqrt{27.440{,}91} = 165{,}65$.

O que se ganha ao converter a variância em seu correspondente desvio padrão? Lembre-se de que as unidades associadas à variância são elevadas ao quadrado. Por exemplo, a variância da amostra dos dados de salários iniciais dos graduados da escola de administração é $s^2 = 27.440{,}91$ (dólares)^2. Uma vez que o desvio padrão é a raiz quadrada da variância, as unidades da variância (dólares elevados ao quadrado) são convertidas em dólares no desvio padrão. Assim, o desvio padrão dos dados de salários iniciais é US$ 165,65. Em outras palavras, o desvio padrão é medido nas mesmas unidades que os dados originais. Por esse motivo, o desvio padrão é mais facilmente comparado à média e a outros dados estatísticos que são medidos nas mesmas unidades que os dados originais.

Coeficiente de Variação

Em algumas situações, podemos estar interessados em uma estatística descritiva que indique qual é o tamanho do desvio padrão em relação à média. Essa medida é chamada **coeficiente de variação** e geralmente é expressa como uma porcentagem.

O coeficiente de variação é uma medida de variabilidade relativa: ele mede o desvio padrão relativo à média.

COEFICIENTE DE VARIAÇÃO

$$\left(\frac{\text{Desvio padrão}}{\text{Média}} \times 100 \right)\% \tag{3.8}$$

Em relação aos dados de tamanhos de classe, descobrimos que a média da amostra é 44 e que o desvio padrão da amostra é 8. O coeficiente de variação é [(8/44) x 100]% = 18,2%. Expressamente, o coeficiente de variação nos diz que o desvio padrão da amostra é 18,2% do valor da média da amostra. Em relação aos dados de salários iniciais com uma média de amostra igual a 2.940 e um desvio padrão da amostra igual a 165,65, o coeficiente de variação, [(165,65/2.940) x 100]% = 5,6%, nos diz que o desvio padrão da amostra é somente 5,6% do valor da média da amostra. Em geral, o coeficiente de variação é uma estatística útil para compararmos a variabilidade de variáveis que têm desvios padrão diferentes e médias diferentes.

NOTAS E COMENTÁRIOS

1. Pacotes de software estatístico e planilhas eletrônicas podem ser usados para desenvolver a estatística descritiva apresentada neste capítulo. Depois que os dados são introduzidos em uma planilha, alguns comandos simples podem ser utilizados para gerar os dados de saída (*output*) desejados. Nos Apêndices 3.1 e 3.2, mostramos como o Minitab e o Excel podem ser usados para desenvolver estatísticas descritivas.
2. O desvio padrão é uma medida usada comumente para se calcular o risco associado ao investimento em ações e fundos de ações (*Business Week*, 17 de janeiro de 2000). Ele fornece uma medida de como os retornos mensais flutuam em torno dos retornos médios de longo prazo.
3. Arredondar o valor da média da amostra $\bar{x}$ e os valores dos desvios elevados ao quadrado $(x_i - \bar{x})^2$ pode levar a erros quando se usa uma calculadora para calcular a variância e o desvio padrão. Para reduzir os erros de arredondamento, recomendamos utilizar pelo menos seis dígitos significativos durante os cálculos intermediários. A variância ou o desvio padrão resultante pode então ser arredondado para uma quantidade menor de dígitos.

4. Uma fórmula alternativa para o cálculo da variância da amostra é

$$s^2 = \frac{\Sigma x_i^2 - n\bar{x}^2}{n - 1}$$

em que $\Sigma x_i^2 = x_1^2 + x_2^2 + \cdots + x_n^2$.

Exercícios

Métodos

13. Considere uma amostra com os valores de dados 10, 20, 12, 17 e 16. Calcule a amplitude e a amplitude interquartil.
14. Considere uma amostra com os valores de dados 10, 20, 12, 17 e 16. Calcule a variância e o desvio padrão.
15. Considere uma amostra com os valores de dados 27, 25, 20, 15, 30, 34, 28 e 25. Calcule a amplitude, a amplitude interquartil, a variância e o desvio padrão.

AUTOTESTE

Aplicações

16. As pontuações de um jogador de boliche em seis jogos foram 182, 168, 184, 190, 170 e 174. Usando esses dados como uma amostra, calcule as seguintes estatísticas descritivas:
    a. Amplitude
    b. Variância
    c. Desvio padrão
    d. Coeficiente de variação

AUTOTESTE

17. Um *home theater* compacto é a maneira mais fácil e mais barata de obter *surround sound* em um centro de diversão doméstico. Uma amostra de preços é apresentada a seguir (*Consumer Reports Buying Guide*, 2004). Os preços referem-se a modelos com DVD *player* e a modelos sem DVD *player*.

Modelos com DVD *Player*	Preço	Modelos sem DVD *Player*	Preço
Sony HT-1800DP	$450	Pioneer HTP-230	$300
Pioneer HTD-330DV	300	Sony HT-DDW750	300
Sony HT-C800DP	400	Kenwood HTB-306	360
Panasonic SC-HT900	500	RCA RT-2600	290
Panasonic SC-MT1	400	Kenwood HTB-206	300

   a. Calcule a média de preços para modelos com DVD *player* e a média de preços para modelos sem DVD *player*. Qual é o preço adicional que se paga para ter um DVD *player* incluído em uma unidade de *home theater*?
   b. Calcule a amplitude, a variância e o desvio padrão das duas amostras. O que essa informação lhe diz a respeito dos preços de modelos com e sem um DVD player?

18. Os preços de aluguel de carro por dia de uma amostra de sete cidades da região leste dos Estados Unidos são os seguintes (*The Wall Street Journal*, 16 de janeiro de 2004):

Cidade	Taxa Diária
Boston	$43
Atlanta	35
Miami	34
Nova York	58
Orlando	30
Pittsburgh	30
Washington, D.C.	36

   a. Calcule a média, a variância e o desvio padrão dos preços de aluguel de carros.
   b. Uma amostra similar de sete cidades da região oeste dos Estados Unidos apresentou um preço médio da amostra correspondente a US$ 38 por dia para o aluguel de carros. A variância e o desvio padrão foram 12,3 e 3,5, respectivamente. Discuta quaisquer diferenças entre os preços de aluguel de carros nas cidades do oeste e do leste dos Estados Unidos.

19. O *Los Angeles Times* publica regularmente um índice da qualidade do ar de várias regiões do sul da Califórnia. Uma amostra dos valores relativos ao índice da qualidade do ar em Pomona forneceu os seguintes dados: 28, 42, 58, 48, 45, 55, 60, 49 e 50.
    a. Calcule a amplitude e a amplitude interquartil.
    b. Calcule a variância da amostra e o desvio padrão da amostra.
    c. Uma amostra de leituras do índice da qualidade do ar em Anaheim forneceu a média da amostra igual a 48,5, uma variância da amostra igual a 136 e o desvio padrão da amostra igual a 11,66. Quais comparações você pode fazer entre a qualidade do ar em Pomona e em Anaheim baseando-se nessas estatísticas descritivas?

20. Os dados apresentados a seguir foram usados para construir os histogramas do número de dias necessários para a Dawson Supply Inc. e a J. C. Clark Distributors emitirem os pedidos de compra (veja a Figura 3.2):

Prazos (Dias) de Entrega da Dawson Supply	11	10	9	10	11	11	10	11	10	10
Prazos (Dias) de Entrega da Clark Distributors	8	10	13	7	10	11	10	7	15	12

Use a amplitude e o desvio padrão para sustentar a observação anterior de que a Dawson Suply apresenta os prazos de entrega mais coerentes e confiáveis.

21. Como os custos dos produtos de mercearia se comparam em todo o território nacional? Usando uma cesta básica de dez itens que incluem farinha de trigo, leite, pão, ovos, café, batatas, cereais e suco de laranja, a revista *Where to Retire* calculou o custo da cesta básica em seis cidades e em seis *retirement areas*[4] de várias partes do território nacional dos Estados Unidos (*Where to Retire*, novembro/dezembro de 2003). Os dados sobre o custo da cesta básica com o menor preço em dólares são os seguintes:

Cidade	Custo	Retirement Area	Custo
Buffalo, NY	$33	Biloxi-Gulfport, MS	$29
Des Moines, IA	27	Asheville, NC	32
Hartford, CT	32	Flagstaff, AZ	32
Los Angeles, CA	38	Hilton Head, SC	34
Miami, FL	36	Fort Myers, FL	34
Pittsburgh, PA	32	Santa Fe, NM	31

a. Calcule a média, a variância e o desvio padrão da amostra de cidades e da amostra de *retirement areas*.
b. Quais observações podem ser feitas com base nas duas amostras?

ARQUIVO DA INTERNET
Broker

22. A American Association of Individual Investors realizou uma pesquisa anual de *discount brokers* (*AAII Journal*, janeiro de 2003). As comissões cobradas pelas 24 *discount brokers* para dois tipos de transações, a comercialização de 100 ações a US$ 50 por ação auxiliada por corretores e a comercialização on-line de 500 ações a US$ 50 por ação, são mostradas na Tabela 3.2.
    a. Calcule a amplitude e a amplitude interquartil de cada tipo de transação.
    b. Calcule a variância e o desvio padrão de cada tipo de transação.
    c. Calcule o coeficiente de variação de cada tipo de transação.
    d. Calcule a variabilidade de custo dos dois tipos de transação.

23. A revista *PC World* publicou avaliações de 15 computadores *notebook* (*PC World*, fevereiro de 2000). Foi utilizada uma escala de 100 pontos para fornecer uma classificação global de cada *notebook*. Uma pontuação na casa dos 90 é excepcional, ao passo que uma pontuação na casa dos 70 é considerada boa. As avaliações globais dos 15 *notebooks* são mostradas a seguir:

[4] NT: Lugar tranqüilo, afastado das grandes cidades, para onde se mudam as pessoas depois de se aposentarem. Lugar de descanso e lazer; retiro.

Notebook	Classsificação Geral
AMS Tech Roadster 15CTA380	67
Compaq Armada M700	78
Compaq Prosignia Notebook 150	79
Dell Inspiron 3700 C466GT	80
Dell Inspiron 7500 R500VT	84
Dell Latitude Cpi A366XT	76
Enpower ENP-313 Pro	77
Gateway Solo 9300LS	92
HP Pavilion Notebook PC	83
IBM ThinkPad I Series 1480	78
Micro Express NP7400	77
Micron TransPort NX PII-400	78
NEC Versa SX	78
Sceptre Soundx 5200	73
Sony VAIO PCG-F340	77

ARQUIVO DA INTERNET
Notebook

Calcule a amplitude, a amplitude interquartil e o desvio padrão dessa amostra de computadores *notebook.*

24. Foram registrados os seguintes tempos pelos corredores de 400 e 1.600 metros de uma equipe de atletismo de uma universidade (os tempos estão expressos em minutos):

Tempos para 400 Metros:	0,92	0,98	1,04	0,90	0,99
Tempos para 1.600 Metros:	4,52	4,35	4,60	4,70	4,50

Depois de ver essa amostra de tempos de corrida, um dos treinadores comentou que os corredores de 400 metros apresentaram tempos mais constantes. Use o desvio padrão e o coeficiente de variação para sintetizar a variabilidade dos dados. O uso do coeficiente de variação indica que a afirmação do treinador se justifica?

3.3 MEDIDAS DA FORMA DA DISTRIBUIÇÃO, DA POSIÇÃO RELATIVA E DETECÇÃO DE PONTOS FORA DA CURVA

Descrevemos diversas medidas de posição e de variabilidade dos dados. Além disso, muitas vezes é importante ter-se a medida da forma de uma distribuição. No Capítulo 2, observamos que um histograma fornece uma apresentação gráfica que mostra a forma de uma distribuição. Uma medida numérica importante da forma de uma distribuição é chamada **assimetria**.

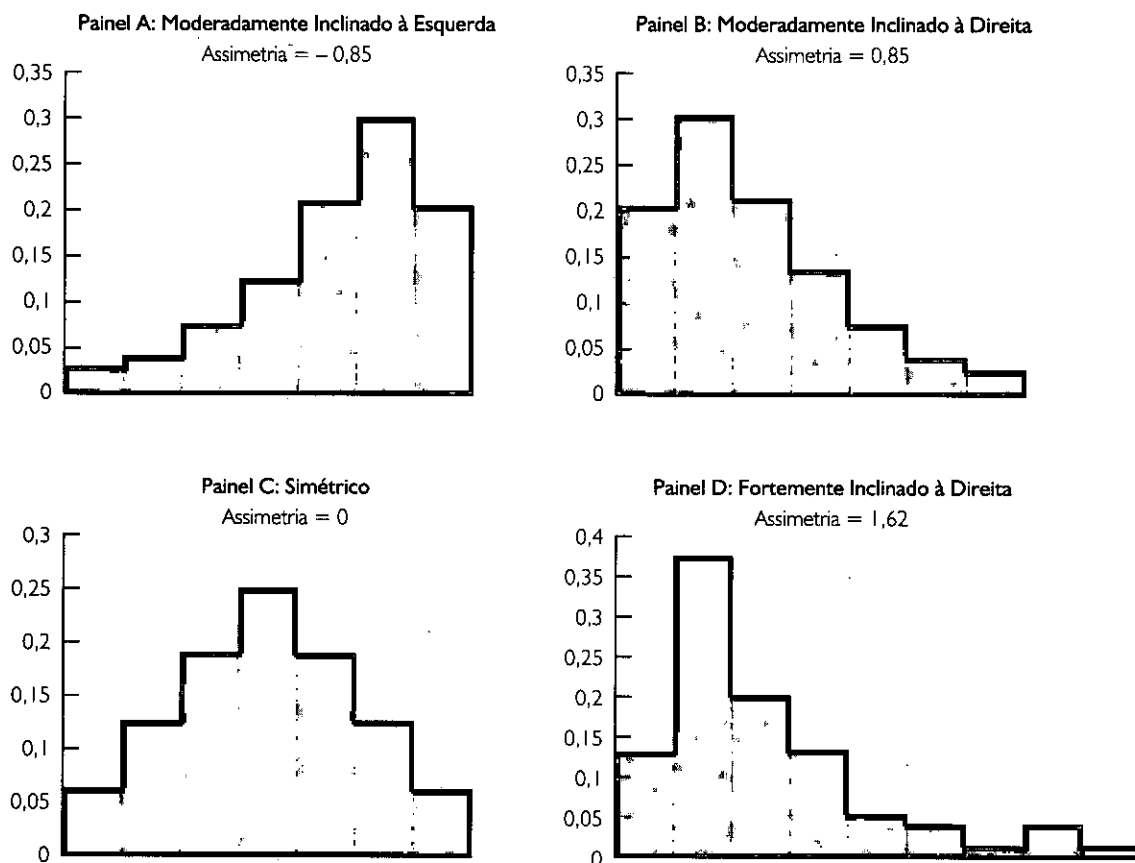
Forma da Distribuição

A Figura 3.3 mostra quatro histogramas construídos a partir de distribuições de freqüência relativa. Os histogramas dos painéis A e B estão moderadamente inclinados. O do painel A está inclinado à esquerda; sua assimetria é de –0,85. O histograma do painel B está inclinado à direita; sua assimetria é de +0,85. O histograma do painel C é simétrico; sua assimetria é nula. O histograma do painel D é fortemente inclinado à direita; sua assimetria é 1,62. A fórmula usada para calcular a assimetria é um tanto complexa.[5] Entretanto, a assimetria pode ser prontamente calculada utilizando-se software estatístico (veja os Apêndices 3.1 e 3.2). Para dados inclinados à esquerda, a assimetria é negativa; para dados inclinados à direita, a assimetria é positiva. Se os dados são simétricos, a assimetria é nula.

Para uma distribuição simétrica, a média e a mediana são iguais. Quando os dados são inclinados positivamente, a média geralmente será maior que a mediana; quando os dados são inclinados negativamente, a média normalmente será menor que a mediana. Os dados utilizados para construir o histograma do painel D são de compras efetuadas por clientes em uma loja de vestuário feminino. A média do valor das compras é US$ 77,60 e a mediana do valor das compras é US$ 59,70. Os relativamente poucos valores de compra elevados tendem a ampliar a média, ao passo que a mediana não é afetada pelos valores de compra elevados. A mediana constitui a medida de posição preferível quando os dados são fortemente assimétricos.

[5] A fórmula para calcular a assimetria de amostras é:

$$\text{Assimetria} = \frac{n}{(n-1)(n-2)} \sum \left(\frac{x_i - \bar{x}}{s} \right)^3$$

Figura 3.3 Histogramas que indicam a assimetria de quatro distribuições

Contagens-z

Além das medidas de posição, de variabilidade e de forma, também estamos interessados na posição relativa dos valores contidos em um conjunto de dados. As medidas de posição relativa nos ajudam a determinar quão afastado um valor em particular está da média.

Usando tanto a média como o desvio padrão, podemos determinar a posição relativa de qualquer observação. Suponha que temos uma amostra de n observações, sendo os valores denotados por $x_1, x_2, \ldots, x_n$. Além disso, suponha que a média da amostra, $\bar{x}$, e o desvio padrão da amostra, s, já tenham sido calculados. Associado a cada valor, x_i, há outro valor que se chama **contagem-z**. A Equação (3.9) mostra como a contagem-z é calculada para cada x_i.

CONTAGEM-z

$$z_i = \frac{x_i - \bar{x}}{s} \tag{3.9}$$

em que

z_i = a contagem-z para x_i
$\bar{x}$= a média da amostra
s = o desvio padrão da amostra

A contagem-z muitas vezes é denominada *valor padronizado*. A contagem-z, z_i, pode ser interpretada como o *número de desvios padrão que x_i está afastado da média $\bar{x}$*. Por exemplo, z_1 = 1,2 indicaria que x_1 é 1,2 desvio padrão maior que a média da amostra. Similarmente, z_2 = –0,5 indicaria que x_2 é 0,5, ou 1/2, desvio padrão menor que a média da amostra. Ocorre uma contagem-z maior que zero para observa-

ções com um valor maior que a média, e ocorre uma contagem-*z* menor que zero para observações com um valor menor que a média. Uma contagem-*z* igual a zero indica que o valor da observação é igual à média.

A contagem-*z* de qualquer observação pode ser interpretada como uma medida da posição relativa da observação no conjunto de dados. Desse modo, pode-se dizer que as observações feitas em dois diferentes conjuntos de dados que possuem a mesma contagem-*z* têm a mesma posição relativa em termos de estarem o mesmo número de desvios padrão afastados da média.

As contagens-*z* dos dados de tamanhos de classe estão calculadas na Tabela 3.5. Lembre-se da média da amostra, $\bar{x} = 44$, e do desvio padrão, $s = 8$, calculados anteriormente. A contagem-*z* de –1,50 correspondente à quinta observação mostra que é a mais afastada da média; ela está 1,50 desvio padrão abaixo da média.

Tabela 3.5 Contagens-*z* dos dados de tamanhos de classe

Número de Estudantes na Classe (x_i)	Desvio em Torno da Média ($x_i - \bar{x}$)	Contagem-z $\left(\frac{x_i - \bar{x}}{s}\right)$
46	2	2/8 = ,25
54	10	10/8 = 1,25
42	−2	22/8 = −,25
46	2	2/8 = ,25
32	−12	212/8 = −1,50

Teorema de Chebyshev

O **teorema de Chebyshev** nos permite fazer afirmações acerca da proporção de valores de dados que devem estar contidos em um número específico de desvios padrão da média.

TEOREMA DE CHEBYSHEV

Pelo menos $(1 - 1/z^2)$ dos valores de dados devem estar contidos em *z* desvios padrão da média, em que *z* é qualquer valor maior que 1.

Algumas das aplicações desse teorema, com $z = 2$, 3 e 4 desvios padrão, são as seguintes:

- Pelo menos 0,75, ou 75%, dos valores de dados devem estar contidos em $z = 2$ desvios padrão da média.
- Pelo menos 0,89, ou 89%, dos valores de dados devem estar contidos em $z = 3$ desvios padrão da média.
- Pelo menos 0,94, ou 94%, dos valores de dados devem estar contidos em $z = 4$ desvios padrão da média.

Como um exemplo do uso do teorema de Chebyshev, suponha que as notas dos exames semestrais de 100 estudantes de um curso de estatística de uma escola de administração tenham obtido a média 70 e um desvio padrão igual a 5. Quantos estudantes tiveram notas de exame entre 60 e 80? Quantos estudantes tiveram notas entre 58 e 82?

Em relação às notas entre 60 e 80, observamos que 60 está dois desvios padrão abaixo da média e que 80 está dois desvios-padrão acima da média. Usando o teorema de Chebyshev, vemos que pelo menos 0,75, ou pelo menos 75%, das observações devem ter valores que estão dentro dos desvios padrão da média. Dessa forma, pelo menos 75% dos estudantes devem ter obtido notas entre 60 e 80.

Em relação às notas entre 58 e 82, vemos que $(58 - 70)/5 = -2,4$ indica que 58 está 2,4 desvios padrão abaixo da média e que $(82 - 70)/5 = +2,4$ indica que 82 está 2,4 desvios padrão acima da média. Aplicando o teorema de Chebyshev com $z = 2,4$, obtemos:

O teorema de Chebyshev requer $z > 1$; mas *z* não precisa ser um número inteiro.

$$\left(1 - \frac{1}{z^2}\right) = \left(1 - \frac{1}{(2,4)^2}\right) = 0,826$$

Pelo menos 82,6% dos estudantes devem ter notas de exame entre 58 e 82.

Regra Empírica

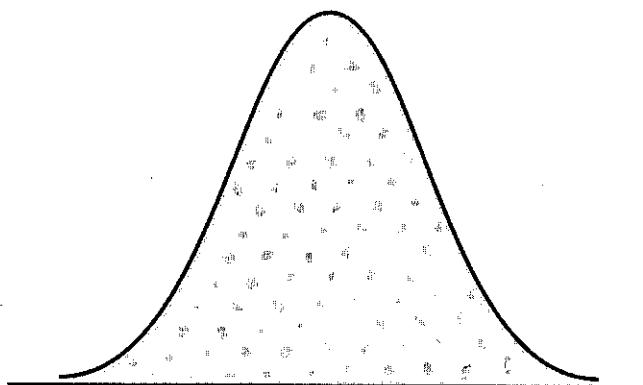
A regra empírica baseia-se na distribuição de probabilidade normal, que será discutida no Capítulo 6. A distribuição normal é extensamente utilizada em todo o livro.

Uma das vantagens do teorema de Chebyshev é que ele se aplica a qualquer conjunto de dados, independentemente da forma da distribuição dos dados. Realmente, ele poderia ser usado com qualquer uma das distribuições da Figura 3.3. Em muitas aplicações práticas, no entanto, os conjuntos de dados exibem uma distribuição simétrica em forma de morro ou de sino, como é mostrado na Figura 3.4. Quando se acredita que os dados se aproximam dessa distribuição, pode-se usar a **regra empírica** para determinar a porcentagem de valores de dados que devem estar contidos em um número específico de desvios padrão da média.

REGRA EMPÍRICA

Para dados que têm uma distribuição em forma de sino:

- Aproximadamente 68% dos valores de dados estarão contidos em um desvio padrão da média.
- Aproximadamente 95% dos valores de dados estarão contidos em dois desvios padrão da média.
- Quase todos os valores de dados estarão contidos em três desvios padrão da média.

Figura 3.4 Uma distribuição simétrica em forma de morro ou sino

Por exemplo, as embalagens de detergente líquido são preenchidas automaticamente em uma linha de produção. Os volumes de preenchimento freqüentemente têm uma distribuição em forma de sino. Se a média dos volumes de preenchimento for 16 onças e o desvio padrão, 0,25 onça, podemos usar a regra empírica para tirar as seguintes conclusões:

- Aproximadamente 68% das embalagens cheias terão cargas entre 15,75 onças e 16,25 onças (dentro de um desvio padrão da média).
- Aproximadamente 95% das embalagens cheias terão cargas entre 15,50 onças e 16,50 onças (dentro de dois desvios padrão da média).
- Quase todas as embalagens cheias terão cargas entre 15,25 onças e 16,75 onças (dentro de três desvios padrão da média).

Detecção de Pontos Fora da Curva

Às vezes, um conjunto de dados terá uma ou mais observações com valores excepcionalmente grandes ou pequenos. Esses valores extremos são chamados **pontos fora da curva**. Estatísticos experientes tomam medidas para identificar os pontos fora da curva e depois revêem cada um deles minuciosamente. Um ponto fora da curva pode ser um valor de dados que foi incorretamente registrado. Se assim for, ele pode ser corrigido antes de prosseguir a análise. Um ponto fora da curva também pode ser proveniente de uma observação que foi incorretamente incluída no conjunto de dados; nesse caso, ela pode ser eliminada. Finalmente, um ponto fora da curva pode ser um valor de dados incomum que foi registrado corretamente e que pertence ao conjunto de dados. Nesses casos, ele deve permanecer.

Valores padronizados (contagens-z) podem ser usados para identificar pontos fora da curva. Lembre-se de que a regra empírica nos permite concluir que, em relação a dados com uma distribuição em forma de sino, quase todos os valores de dados estarão contidos em três desvios padrão da média. Portanto, ao usar contagens-z para identificar pontos fora da curva, recomendamos tratar quaisquer valores de dados com uma contagem-z menor que –3 ou maior que +3 como um ponto fora da curva. Esses valores de dados podem então ser revisados quanto à precisão e para determinar se pertencem ao conjunto de dados.

Consulte as contagens-z referentes aos dados de tamanhos de classe da Tabela 3.5. A contagem-z igual a –1,50 mostra que o quinto tamanho de classe é o mais afastado da média. Entretanto, esse valor padronizado está bem dentro da diretriz –3 a +3 para pontos fora da curva. Desse modo, as contagens-z não indicam se há pontos fora da curva nos dados de tamanho de classe.

É uma boa idéia verificar se há pontos fora da curva antes de tomar decisões baseadas em análise de dados. Freqüentemente se cometem erros ao fazer o registro de dados e ao digitá-los no computador. Pontos fora da curva não devem ser necessariamente excluídos, mas sua precisão e adequabilidade devem ser verificadas.

NOTAS E COMENTÁRIOS

1. O teorema de Chebyshev é aplicável a qualquer conjunto de dados e pode ser usado para estabelecer o número mínimo de valores de dados que estarão dentro de certo número de desvios padrão da média. Quando se sabe que os dados têm aproximadamente a forma de sino, pode-se dizer mais coisas. Por exemplo, a regra empírica nos permite dizer que *aproximadamente* 95% dos valores de dados estarão dentro de dois desvios padrão da média; o teorema de Chebyshev nos permite concluir somente que pelo menos 75% dos valores de dados estarão nesse intervalo.
2. Antes de analisar um conjunto de dados, os estatísticos geralmente fazem uma série de verificações para assegurar a validade dos dados. Em um estudo de grande porte não é incomum a ocorrência de erros ao registrar valores de dados ou ao digitá-los no computador. Identificar pontos fora da curva é uma ferramenta utilizada para conferir a validade dos dados.

Exercícios

Métodos

25. Considere uma amostra com os valores de dados 10, 20, 12, 17 e 16. Calcule a contagem-z de cada uma das cinco observações.
26. Considere uma amostra com a média 500 e desvio padrão 100. Quais são as contagens-z dos seguintes valores de dados: 520, 650, 500, 450 e 280?

AUTOTESTE

27. Considere uma amostra com a média 30 e desvio padrão 5. Use o teorema de Chebyshev para determinar a porcentagem dos dados que se encontram dentro de cada uma das seguintes amplitudes:
    a. 20 a 40
    b. 15 a 45
    c. 22 a 38
    d. 18 a 42
    e. 12 a 48
28. Suponha que os dados tenham uma distribuição em forma de sino com uma média igual a 30 e desvio padrão, 5. Use a regra empírica para determinar a porcentagem de dados que se encontram dentro de cada uma das seguintes amplitudes:
    a. 20 a 40
    b. 15 a 45
    c. 25 a 35

Aplicações

AUTOTESTE

29. Os resultados de uma pesquisa em nível nacional mostraram que, em média, os adultos dormem 6,9 horas por noite (2000 Omnibus Sleep in America Poll). Suponha que o desvio padrão seja de 1,2 hora.
    a. Use o teorema de Chebyshev para calcular a porcentagem de indivíduos que dormem entre 4,5 e 9,3 horas.
    b. Use o teorema de Chebyshev para calcular a porcentagem de indivíduos que dormem entre 3,9 e 9,9 horas.

c. Suponha que o número de horas de sono segue uma distribuição em forma de sino. Use a regra empírica para calcular a porcentagem de indivíduos que dormem entre 4,5 e 9,3 horas por dia. Como esse resultado se compara com o valor que você obteve ao usar o teorema de Chebyshev do item (a)?

30. A Energy Information Administration publicou que o preço médio de varejo por galão de gasolina comum era US$ 1,47 (*The Wall Street Journal*, 30 de janeiro de 2003). Suponha que o desvio padrão tenha sido US$ 0,08 e que o preço de varejo por galão tenha uma distribuição em forma de sino.

    a. Qual porcentagem de gasolina comum foi vendida entre US$ 1,39 e US$ 1,55 por galão?
    b. Qual porcentagem de gasolina comum foi vendida entre US$ 1,39 e US$ 1,63 por galão?
    c. Qual porcentagem de gasolina comum foi vendida a mais de US$ 1,63 e por galão?

31. A média nacional para a parte oral do College Board's Scholastic Aptitude Test (SAT)[6] é 507 (*The World Almanac*, 2004). O College Board reescalona periodicamente as notas do exame de tal forma que o desvio padrão seja aproximadamente 100. Responda às perguntas a seguir usando uma distribuição em forma de sino e a regra empírica para as notas do exame oral.

    a. Qual é a porcentagem dos estudantes que têm notas superiores a 607 no exame oral do SAT?
    b. Qual é a porcentagem dos estudantes que notas superiores a 707 no exame oral do SAT?
    c. Qual é a porcentagem dos estudantes que têm notas entre 407 e 507 no exame oral do SAT?
    d. Qual é a porcentagem dos estudantes que têm notas entre 307 e 607 no exame oral do SAT?

32. Os elevados custos praticados no mercado imobiliário da Califórnia fizeram que as famílias que não possam se dar ao luxo de comprar casas maiores considerem as construções de quintal como uma forma alternativa de expandir suas residências. Muitas utilizam as estruturas existentes em seus quintais como escritórios, estúdios artísticos e áreas de lazer, bem como para armazenamento adicional. O preço médio de uma construção de quintal personalizada, feita em madeira e coberta com telhas de amianto é US$ 3.100 (*Newsweek*, 29 de setembro de 2003). Suponha que o desvio padrão seja US$ 1.200.

    a. Qual é a contagem-z de uma estrutura de quintal que custa US$ 2.300?
    b. Qual é a contagem-z de uma estrutura de quintal que custa US$ 4.900?
    c. Interprete a contagem-z dos itens (a) e (b). Comente se um deles seria considerado um ponto fora da curva.-
    d. O artigo da *Newsweek* descreveu a combinação de uma edícula-escritório construída em Albany, Califórnia, por US$ 13 mil. Essa estrutura deveria ser considerada um ponto fora da curva? Explique.

33. A Wageweb realiza pesquisas de dados salariais e apresenta sumários em seu site. Os salários registrados para gerentes de benefícios variam de US$ 50.935 a US$ 79.577 (Wageweb.com, 12 de abril de 2000). Suponha que os dados a seguir sejam uma amostra dos salários anuais de 30 gerentes de benefícios. Os dados estão expressos em milhares de dólares:

57,7	64,4	62,1	59,1	71,1
63,0	64,7	61,2	66,8	61,8
64,2	63,3	62,2	61,2	59,4
63,0	66,7	60,3	74,0	62,8
68,7	63,.8	59,2	60,3	56,6
59,3	69,5	61,7	58,9	63,1

ARQUIVO DA INTERNET
WageWeb

    a. Calcule a média e o desvio padrão dos dados da amostra.
    b. Usando a média e o desvio padrão calculados no item (a) como estimativas da média e do desvio padrão dos salários da população de gerentes de benefícios, use o teorema de Chebyshev para determinar a porcentagem de gerentes de benefícios que têm salários anuais entre US$ 55 mil e US$ 71mil.

[6] NT: Exame promovido pelas universidades norte-americanas como parte do processo de seleção de estudantes para admissão ao curso superior; ele é realizado sete vezes por ano, envolvendo matemática e inglês. Há sete seções: três de matemática, três orais, e uma prática (experimental) que não recebe notas, mas é usada somente para pesquisa.

c. Desenvolva um histograma dos dados da amostra. Um software de computador fornece 0,97 como medida de assimetria. Parece razoável supor que a distribuição de salários anuais possa ser aproximada por uma distribuição em forma de sino?

d. Suponha que a distribuição de salários anuais tenha a forma de sino. Usando a média e o desvio padrão computados no item (a) como estimativa da média e do desvio padrão dos salários da população de gerentes de benefícios, use a regra empírica para determinar a porcentagem de gerentes de benefícios que têm salários anuais entre US$ 55 mil e US$ 71 mil Compare sua resposta com o valor calculado no item (b).

e. Os dados amostrais contêm algum dado fora da curva?

34. Uma amostra de 10 pontuações de jogos de basquete universitário da NCAA[7] forneceu os seguintes dados (*USA Today*, 26 de janeiro de 2004).

Time Vencedor	Pontos	Time Perdedor	Pontos	Margem de Vitórias
Arizona	90	Oregon	66	24
Duke	85	Georgetown	66	19
Florida State	75	Wake Forest	70	5
Kansas	78	Colorado	57	21
Kentucky	71	Notre Dame	63	8
Louisville	65	Tennessee	62	3
Oklahoma State	72	Texas	66	6
Purdue	76	Michigan State	70	6
Stanford	77	Southern Cal	67	10
Wisconsin	76	Illinois	56	20

ARQUIVO DA INTERNET
NCAA

a. Calcule a média e o desvio padrão dos pontos marcados pelo time vencedor.

b. Suponha que os pontos marcados pelo time vencedor em todos os jogos da NCAA sigam uma distribuição em forma de sino. Usando a média e o desvio padrão encontrados no item (a), estime a porcentagem de todos os jogos da NCAA em que o time vencedor obtém 84 ou mais pontos. Estabeleça a porcentagem de jogos da NCAA em que o time vencedor marca mais de 90 pontos.

c. Calcule a média e o desvio padrão da margem de vitórias. Os dados contêm pontos fora da curva? Explique.

35. A *Consumer Review* publica análises do desempenho e avaliações da qualidade de uma série de produtos na internet. Os dados a seguir são uma amostra de 20 sistemas de alto-falantes e suas respectivas avaliações (http://www.audioreview.com). As avaliações são apresentadas em uma escala de 1 a 5, sendo 5 a melhor.

Alto-falante	Avaliação	Alto-falante	Avaliação
Infinity Kappa 6.1	4,00	ACI Sapphire III	4,67
Allison One	4,12	Bose 501 Series	2,14
Cambridge Ensemble II	3,82	DCM KX-212	4,09
Dynaudio Contour 1.3	4,00	Eosone RSF1000	4,17
Hsu Rsch. HRSW12V	4,56	Joseph Audio RM7si	4,88
Legacy Audio Focus	4,32	Martin Logan Aerius	4,26
Mission 73li	4,33	Omni Audio SA 12.3	2,32
PSB 400i	4,50	Polk Audio RT12	4,50
Snell Acoustics D IV	4,64	Sunfire True Subwoofer	4,17
Thiel CS1.5	4,20	Yamaha NS-A636	2,17

ARQUIVO DA INTERNET
Speakers

a. Calcule a média e a mediana.

b. Calcule o primeiro e o terceiro quartis.

c. Calcule o desvio padrão.

d. A assimetria desses dados é –1,67. Comente a forma da distribuição.

e. Quais são as contagens-z associadas à Allison One e à Omni Audio?

f. Os dados contêm algum ponto fora da curva? Explique.

[7] NT: National Collegiate Athletic Association.

3.4 ANÁLISE EXPLORATÓRIA DE DADOS

No Capítulo 2, introduzimos a apresentação de ramo-e-folha como uma técnica de análise exploratória dos dados. Lembre-se de que a análise exploratória dos dados nos permite usar cálculos aritméticos simples e gráficos fáceis de desenhar para sintetizar os dados. Nesta seção, prosseguiremos a análise exploratória de dados considerando a regra de cinco itens e desenhos esquemáticos (*box plots*).

Regra de Cinco Itens

Em uma **regra de cinco itens**, os cinco números seguintes são usados para sintetizar os dados:

1. Menor valor
2. Primeiro quartil (Q_1)
3. Mediana (Q_2)
4. Terceiro quartil (Q_3)
5. Maior valor

A maneira mais fácil de desenvolver uma regra de cinco itens é colocar primeiramente os dados em ordem crescente. Depois é fácil identificar o menor valor, os três quartis e o maior valor. Os salários mensais iniciais mostrados na Tabela 3.1 correspondentes a uma amostra de 12 graduados da escola de administração são repetidos aqui em ordem crescente:

2.710 2.755 2.850 | 2.880 2.880 2.890 | 2.920 2.940 2.950 | 3.050 3.130 3.325

Q_1 = 2.865 Q_2 = 2.905 Q_3 = 3.000

Mediana

A mediana de 2.905 e os quartis Q_1 = 2.865 e Q_2 = 3.000 foram calculados na Seção 3.1. Uma revisão dos dados nos mostra que o menor valor é 2.710 e o maior valor é 3.325. Desse modo, a regra de cinco itens correspondente aos dados salariais é 2.710, 2.865, 2.905, 3.000 e 3.325. Aproximadamente um quarto, ou 25%, das observações se encontram entre números adjacentes em uma regra de cinco itens.

Os desenhos esquemáticos (*box plots*) constituem outra maneira de se identificar pontos fora da curva. Mas eles não identificam necessariamente os mesmos valores, por exemplo, aqueles que têm uma contagem-z menor que −3 ou maior que +3. Tanto o primeiro como o segundo, ou ambos os procedimentos, podem ser usados.

Desenhos Esquemáticos (*Box Plots*)

Um **desenho esquemático** é um sumário gráfico de dados que se baseia em uma regra de cinco itens. A chave para o desenvolvimento de um desenho esquemático é o cálculo da mediana e dos quartis, Q_1 e Q_2. A amplitude interquartil, AIQ = $Q_3 - Q_1$, também é usada. A Figura 3.5 representa o desenho esquemático dos dados de salários mensais iniciais. Os passos para construirmos o desenho esquemático são os seguintes:

1. Desenhamos um retângulo em que suas extremidades se localizam no primeiro e terceiro quartis. Em relação aos dados salariais, Q_1 = 2.865 e Q_3 = 3.000. Esse retângulo contém os 50% intermediários dos dados.
2. Desenhamos uma linha vertical no retângulo, na posição da mediana (2.905 para os dados salariais).
3. Ao usar a amplitude interquartil, AIQ = $Q_3 - Q_1$, localizamos os *limites*. Os limites do desenho esquemático estão 1,5(AIQ) abaixo de Q_1 e 1,5(AIQ) acima de Q_3. Em relação aos dados salariais, AIQ = $Q_3 - Q_1$ = 3.000 − 2.865 = 135. Desse modo, os limites são 2.865 − 1,5(135) = 2.662,5 e 3.000 + 1,5(135) = 3.202,5. Os dados fora desses limites são considerados *dados fora da curva*.
4. As linhas tracejadas da Figura 3.5 são chamadas *costeletas*. As costeletas são desenhadas das bordas do retângulo até os valores mínimo e máximo localizados *dentro dos limites* calculados na etapa 3. Assim, as costeletas terminam nos valores salariais de 2.710 e 3.130.
5. Finalmente, a posição de cada ponto fora da curva é indicada pelo símbolo *. Na Figura 3.5, vemos um ponto fora da curva: 3.325.

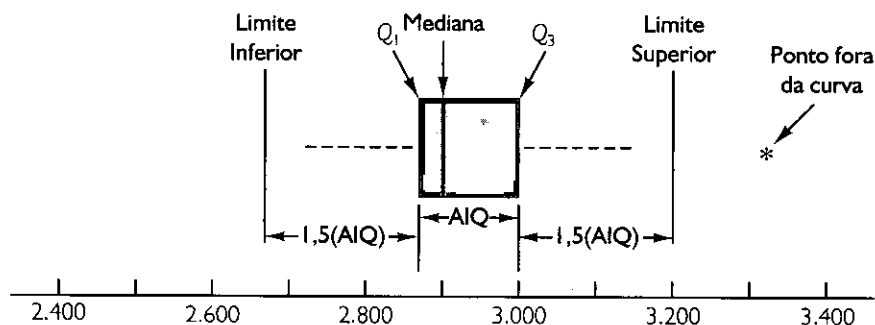
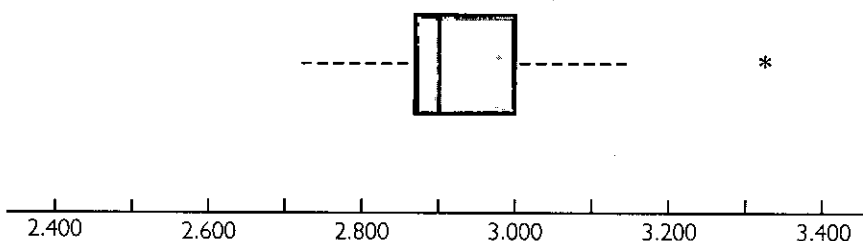
Figura 3.5 Desenho esquemático (*box plot*) dos dados de salários iniciais com linhas indicando os limites mínimo e máximo

Na Figura 3.5 incluímos linhas que indicam a posição dos limites superior e inferior. Essas linhas foram traçadas para indicar como os limites são calculados e onde eles se localizam em relação aos dados salariais. Não obstante os limites sempre serem calculados, geralmente eles não são traçados nos desenhos esquemáticos. A Figura 3.6 mostra a aparência habitual de um desenho esquemático (*box plot*) correspondente aos dados salariais.

Figura 3.6 Desenho esquemático (*box plot*) dos dados de salários iniciais

NOTAS E COMENTÁRIOS

1. Uma vantagem da análise exploratória de dados é que elas são fáceis de usar; poucos cálculos numéricos são necessários. Simplesmente classificamos os valores de dados em ordem crescente e identificamos a regra de cinco itens. O desenho esquemático (*box plot*) pode ser construído. Não é necessário calcular a média e o desvio padrão dos dados.
2. No Apêndice 3.1 mostramos como construir um desenho esquemático dos dados de salários iniciais usando o Minitab. O desenho esquemático obtido se assemelha exatamente ao da Figura 3.6, mas com um giro de 90º no sentido anti-horário.

Exercícios

Métodos

36. Considere uma amostra com os valores de dados 27, 25, 20, 15, 30, 34, 28 e 25. Apresente a regra de cinco itens dos dados.

AUTOTESTE

37. Apresente o desenho esquemático dos dados do Exercício 36.
38. Apresente a regra de cinco itens e o desenho esquemático dos seguintes dados: 5, 15, 18, 10, 8, 12, 16, 10, 6.
39. Um conjunto de dados tem o primeiro quartil igual a 42 e o terceiro quartil igual a 50. Calcule os limites mínimo e máximo do desenho esquemático correspondente. Um valor de dados igual a 65 deve ser considerado um ponto fora da curva?

Aplicações

40. A Ebby Halliday Realtors faz anúncios publicitários de propriedades e imóveis de primeira classe localizados em todo o território dos Estados Unidos. Os preços relacionados para 12 propriedades e imóveis de primeira classe são mostrados a seguir (*The Wall Street Journal*, 16 de janeiro de 2004). Os preços estão expressos em milhares de dólares:

ARQUIVO DA INTERNET
Property

1.500	700	2.995
895	619	880
719	725	3.100
619	739	1.699
625	799	1.120
4.450	2.495	1.250
2.200	1.395	912
1.280		

a. Apresente uma regra de cinco itens.
b. Calcule os limites inferior e superior.
c. A propriedade que tem o preço mais elevado, US$ 4.450 mil, está relacionada como um imóvel que tem vista panorâmica do White Rock Lake, em Dallas. Essa propriedade deve ser considerada um ponto fora da curva? Explique.
d. A segunda propriedade com preço mais elevado, relacionado como US$ 3.100.000 deve ser considerada um ponto fora da curva? Explique.
e. Apresente um desenho esquemático (*box plot*).

41. As vendas anuais, em milhões de dólares, de 21 empresas produtoras de produtos farmacêuticos são apresentadas a seguir:

AUTOTESTE

8.408	1.374	1.872	8.879	2.459	11.413
608	14.138	6.452	1.850	2.818	1.356
10.498	7.478	4.019	4.341	739	2.127
3.653	5.794	8.305			

a. Apresente uma regra de cinco itens.
b. Calcule os limites inferior e superior.
c. Os dados contêm algum ponto fora da curva?
d. As vendas da Johnson & Johnson são as maiores da lista, com US$ 14.138 milhões. Suponha ter havido um erro de lançamento (uma transposição) e que as vendas foram de US$ 41.138 milhões. O método de detecção de pontos fora da curva do item (c) identifica o problema e possibilita a correção do erro de lançamento?
e. Apresente um desenho esquemático (*box plot*).

42. As folhas de pagamento dos times da Major League Baseball continuam a crescer. As folhas de pagamento, em milhões, são apresentadas a seguir (*The Miami Herald*, 22 de maio de 2002):

ARQUIVO DA INTERNET
Payroll

Time	Folha de Pagamento	Time	Folha de Pagamento
Anaheim	$ 62	Milwaukee	$ 50
Arizona	103	Minnesota	40
Atlanta	93	Montreal	39
Baltimore	60	NY Mets	95
Boston	108	NY Yankees	126
Chi Cubs	76	Oakland	40
Chi White Sox	57	Philadelphia	58
Cincinnati	45	Pittsburgh	42
Cleveland	79	San Diego	41
Colorado	57	San Francisco	78
Detroit	55	Seattle	90
Florida	42	St. Louis	74
Houston	63	Tampa Bay	34
Kansas City	47	Texas	105
Los Angeles	95	Toronto	77

a. Qual é a mediana da folha de pagamento dos times?
b. Apresente uma regra de cinco itens.
c. A folha de pagamento de US$ 126 dos New York Yankees é um ponto fora da curva? Explique.
d. Apresente um desenho esquemático (*box plot*).

43. O presidente do conselho administrativo da Bolsa de Valores de Nova York (*New York Stock Exchange* – Nyse), Richard Grasso, e o conselho de diretores sofreram severas críticas em decorrência do conjunto das remunerações pagas a Grasso. No que se refere a salário, mais bonificações, os US$ 8,5 milhões pagos a Grasso superavam demasiadamente o que ganhavam os altos executivos de todas as principais instituições de serviços financeiros. Os dados a seguir mostram o salário anual total, mais bonificações, pago à alta gerência de 14 instituições de serviços financeiros (*The Wall Street Journal*, 17 de setembro de 2003). Os dados estão expressos em milhões de dólares:

Empresa	Salário/Bonificação	Empresa	Salário/Bonificação
Aetna	$3,5	Fannie Mae	$4,3
AIG	6,0	Federal Home Loan	0,8
Allstate	4,1	Fleet Boston	1,0
American Express	3,8	Freddie Mac	1,2
Chubb	2,1	Mellon Financial	2,0
Cigna	1,0	Merrill Lynch	7,7
Citigroup	1,0	Wells Fargo	8,0

a. Qual é a mediana dos salários anuais, mais bonificações, paga à alta gerência das 14 instituições de serviços financeiros?
b. Apresente uma regra de cinco itens.
c. O salário anual, mais bonificações, pago a Grasso deve ser considerado um ponto fora da curva para esse grupo de altos executivos? Explique.
d. Apresente um desenho esquemático.

ARQUIVO DA INTERNET
Mutual

44. Uma relação de 46 fundos mútuos e suas respectivas porcentagens de rentabilidade total em 12 meses são apresentadas na Tabela 3.6 (*Smart Money*, fevereiro de 2004).

a. Qual é a média e a mediana das porcentagens de rentabilidade desses fundos mútuos?
b. Quais são o primeiro e o terceiro quartis?
c. Apresente uma regra de cinco itens.
d. Os dados contêm algum ponto fora da curva? Apresente um desenho esquemático (*box plot*).

Tabela 3.6 Rentabilidade de fundos mútuos em 12 meses

Fundo Mútuo	Rentabilidade (%)	Fundo Mútuo	Rentabilidade (%)
Alger Capital Appreciation	23,5	Nations Small Company	21,4
Alger LargeCap Growth	22,8	Nations SmallCap Index	24,5
Alger MidCap Growth	38,3	Nations Strategic Growth	10,4
Alger SmallCap	41,3	Nations Value Inv	10,8
AllianceBernstein Technology	40,6	One Group Diversified Equity	10,0
Federated American Leaders	15,6	One Group Diversified Int'l	10,9
Federated Capital Appreciation	12,4	One Group Diversified Mid Cap	15,1
Federated Equity-Income	11,5	One Group Equity Income	6,6
Federated Kaufmann	33,3	One Group Int'l Equity Index	13,2
Federated Max-Cap Index	16,0	One Group Large Cap Growth	13,6
Federated Stock	16,9	One Group Large Cap Value	12,8
Janus Adviser Int'l Growth	10,3	One Group Mid Cap Growth	18,7
Janus Adviser Worldwide	3,4	One Group Mid Cap Value	11,4
Janus Enterprise	24,2	One Group Small Cap Growth	23,6
Janus High-Yield	12,1	PBHG Growth	27,3
Janus Mercury	20,6	Putnam Europe Equity	20,4
Janus Overseas	11,9	Putnam Int'l Capital Opportunity	36,6
Janus Worldwide	4,1	Putnam International Equity	21,5
Nations Convertible Securities	13,6	Putnam Int'l New Opportunity	26,3
Nations Int'l Equity	10,7	Strong Advisor Mid Cap Growth	23,7
Nations LargeCap Enhd. Core	13,2	Strong Growth 20	11,7
Nations LargeCap Index	13,5	Strong Growth Inv	23,2
Nation MidCap Index	19,5	Strong Large Cap Growth	14,5

3.5 MEDIDAS DE ASSOCIAÇÃO ENTRE DUAS VARIÁVEIS

Até aqui, examinamos os métodos numéricos utilizados para sintetizar dados correspondentes a *uma variável a cada vez*. Freqüentemente um gerente ou tomador de decisões está interessado na *relação entre duas variáveis*. Nesta seção apresentamos a covariância e a correlação como medidas descritivas da relação entre duas variáveis.

Iniciamos reconsiderando a aplicação que diz respeito a uma loja de equipamentos de som localizada em São Francisco, conforme apresentamos na Seção 2.4. O gerente da loja quer determinar a relação entre o número de comerciais de televisão divulgados nos fins de semana e as vendas na loja durante a semana seguinte. Dados de amostra com as vendas expressas em centenas de dólares são apresentados na Tabela 3.7. Ela apresenta 10 observações (n = 10), sendo uma para cada semana. O diagrama de dispersão da Figura 3.7 exibe uma relação positiva, com vendas mais elevadas (y) associadas a um número maior de comerciais (x). Realmente, o diagrama de dispersão sugere que uma linha reta poderia ser usada como uma aproximação da relação. Na discussão a seguir, introduzimos a **covariância** como uma medida descritiva da associação linear entre duas variáveis.

Covariância

Para uma amostra de tamanho n com as observações (x_1, y_1), (x_2, y_2) etc., a covariância da amostra é definida da seguinte maneira:

COVARIÂNCIA DA AMOSTRA

$$s_{xy} = \frac{\Sigma(x_i - \bar{x})(y_i - \bar{y})}{n - 1} \qquad \textbf{(3.10)}$$

Essa fórmula emparelha cada x_i com um y_i. Somamos então os produtos obtidos ao multiplicarmos o desvio que cada x_i tem de sua média da amostra $\bar{x}$ pelo desvio que o y_i correspondente tem de sua média da amostra; essa soma é então dividida por $n - 1$.

ARQUIVO DA INTERNET
Stereo

Tabela 3.7 Dados de amostra referentes à loja de equipamentos de som

Semana	Número de Comerciais x	Volume de Vendas ($100s) y
1	2	50
2	5	57
3	1	41
4	3	54
5	4	54
6	1	38
7	5	63
8	3	48
9	4	59
10	2	46

Para medir a intensidade da relação linear entre o número de comerciais x e o volume de vendas y no problema da loja de equipamentos de som, usamos a Equação (3.10) para calcular a covariância da amostra. Os cálculos da Tabela 3.8 apresentam o cálculo de $\Sigma(x_i - \bar{x})(y_i - \bar{y})$. Note que $\bar{x} = 30/10 = 3$ e $\bar{y} = 510/10 = 51$. Usando a Equação (3.10), obtemos a covariância da amostra:

$$s_{xy} = \frac{\Sigma(x_i - \bar{x})(y_i - \bar{y})}{n - 1} = \frac{99}{9} = 11$$

Figura 3.7 Diagrama de dispersão da loja de equipamentos de som

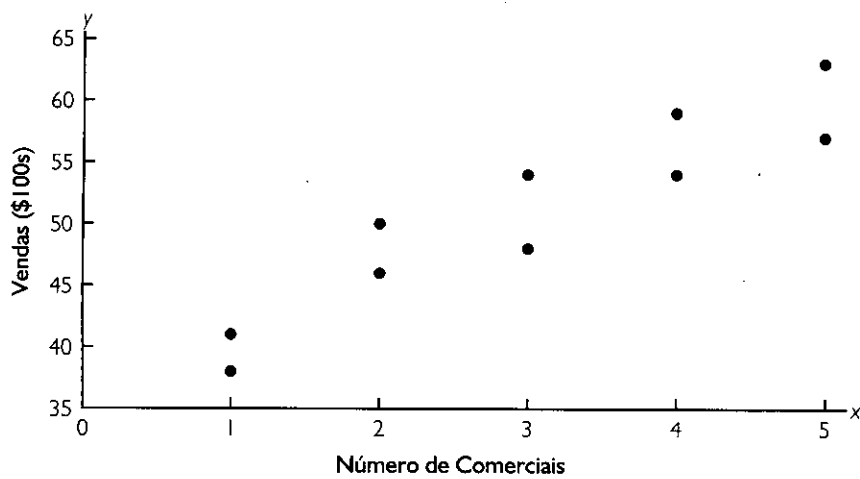



Tabela 3.8 Cálculos da covariância da amostra

	x_i	y_i	$x_i - \bar{x}$	$y_i - \bar{y}$	$(x_i - \bar{x})(y_i - \bar{y})$
	2	50	−1	−1	1
	5	57	2	6	12
	1	41	−2	−10	20
	3	54	0	3	0
	4	54	1	3	3
	1	38	−2	−13	26
	5	63	2	12	24
	3	48	0	−3	0
	4	59	1	8	8
	2	46	−1	−5	5
Totais	30	510	0	0	99

$$s_{xy} = \frac{\Sigma(x_i - \bar{x})(y_i - \bar{y})}{n - 1} = \frac{99}{10 - 1} = 11$$

A fórmula para computar a covariância de uma população de tamanho N é similar à Equação 3.10, mas usamos uma notação diferente para indicar que estamos trabalhando com a população inteira.

COVARIÂNCIA POPULACIONAL

$$\sigma_{xy} = \frac{\Sigma(x_i - \mu_x)(y_i - \mu_y)}{N} \qquad (3.11)$$

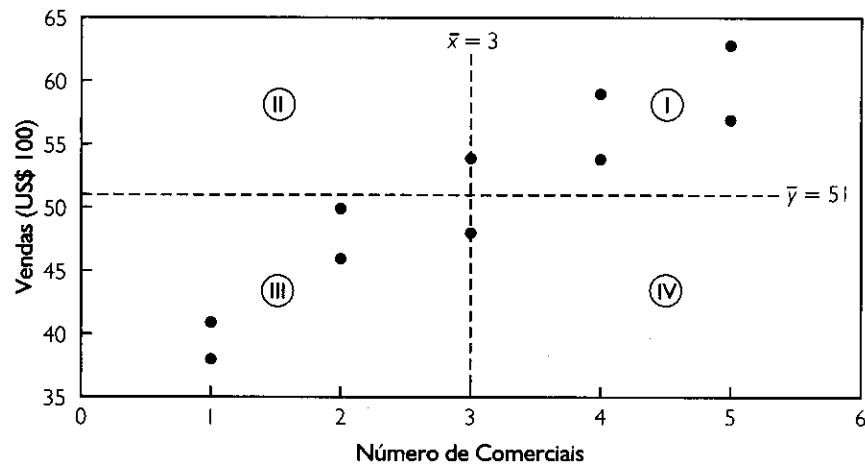
Na Equação 3.11, utilizamos a notação m_x para a média da população da variável x, e μ_y para a média da população da variável y. A covariância populacional σ_{xy} é definida para uma população de tamanho N.

Interpretação da Covariância

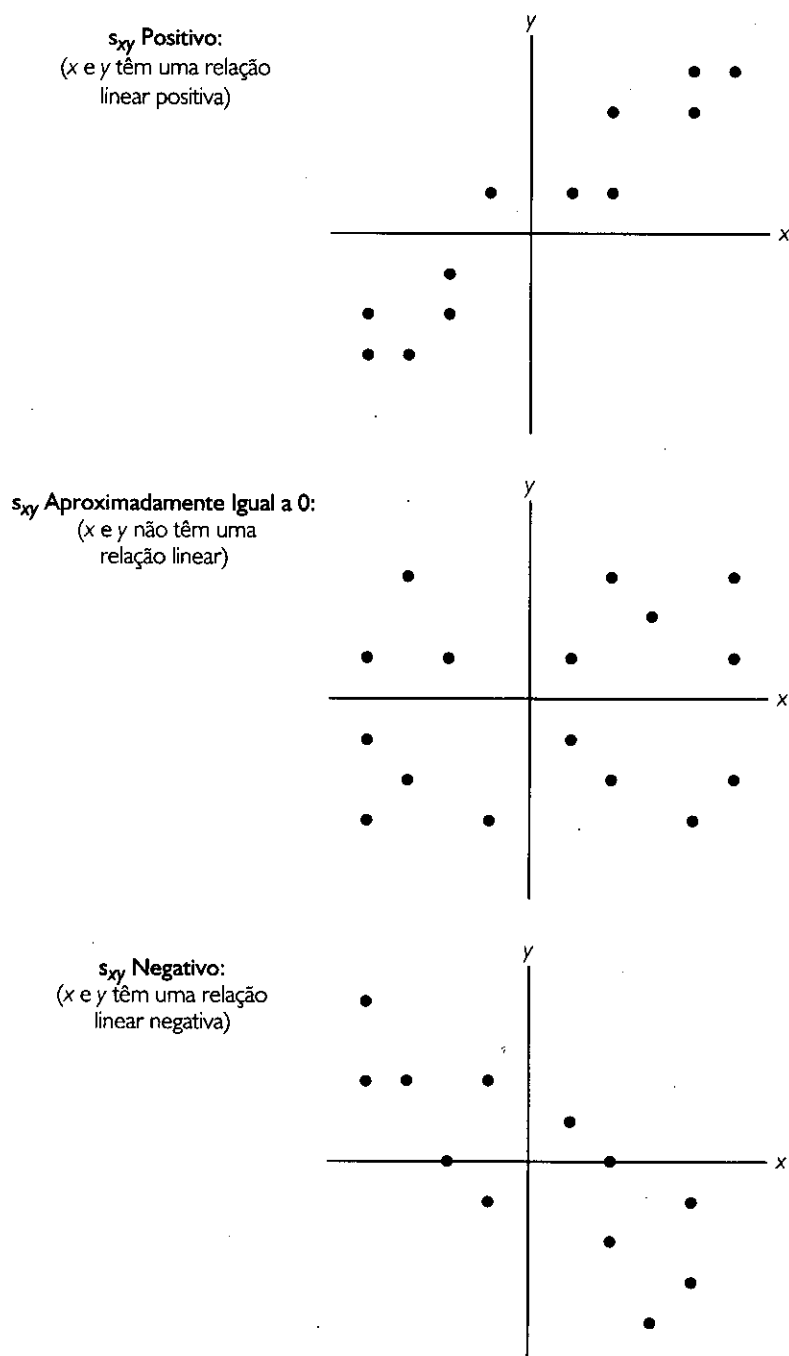
Para auxiliar na interpretação da covariância da amostra, considere a Figura 3.8. Ela é idêntica ao diagrama de dispersão da Figura 3.7, com uma linha tracejada vertical em $\bar{x} = 3$ e uma linha tracejada horizontal em $\bar{y} = 51$. As linhas dividem o gráfico em quatro quadrantes. Os pontos localizados no quadrante I correspondem a x_i maior que $\bar{x}$ e y_i maior que $\bar{y}$, os pontos localizados no quadrante II referem-se a x_i menor que $\bar{x}$ e y_i maior que $\bar{y}$ e assim por diante. Desse modo, o valor $(x_i - \bar{x})$ $(y_i - \bar{y})$ deve ser positivo para pontos localizados no quadrante I, negativo para pontos localizados no quadrante II, positivo para pontos localizados no quadrante II e negativo para pontos localizados no quadrante IV.

Se o valor de s_{xy} for positivo, os pontos que têm a maior influência sobre s_{xy} devem estar nos quadrantes I e III. Portanto, um valor positivo para s_{xy} indica uma associação linear positiva entre x e y; ou seja, à medida que o valor de x se expande, o valor de y aumenta. Se, entretanto, o valor de s_{xy} for negativo, os pontos que têm a maior influência sobre s_{xy} estão nos quadrantes II e IV. Portanto, um valor negativo para s_{xy} indica uma associação linear negativa entre x e y; ou seja, à medida que o valor de x aumenta, o valor de y diminui. Finalmente, se os pontos estiverem uniformemente distribuídos em todos os quatro quadrantes, o valor de s_{xy} se aproximará de zero, indicando que não há nenhuma associação linear entre x e y. A Figura 3.9 apresenta os valores de s_{xy} que se pode esperar com três diferentes tipos de diagramas de dispersão.

A covariância é uma medida da associação linear entre duas variáveis.

Figura 3.8 Diagrama de dispersão da loja de equipamentos de som dividido em quadrantes

Consultando novamente a Figura 3.8, observamos que o diagrama de dispersão da loja de equipamentos de som segue o padrão apresentado no painel da parte superior da Figura 3.9. Como se poderia esperar, o valor da covariância da amostra indica uma relação linear positiva com $s_{xy} = 11$.

Pelo que foi exposto na discussão anterior, poderia parecer que um valor positivo elevado para a covariância indicaria uma relação linear positiva forte e que um valor negativo elevado apontaria uma relação linear negativa forte. Entretanto, um problema quando se usa a covariância como uma medida da intensidade da relação linear é que o valor da covariância depende das unidades de medida para x e y. Por exemplo, suponha que estejamos interessados na relação entre a altura x e o peso y das pessoas. Evidentemente, a intensidade da relação deve ser a mesma se medirmos a altura em centímetros ou em polegadas. Porém, quando a altura é medida em polegadas, obtemos valores numéricos muito mais elevados para $(x_i - \bar{x})$ do que quando medimos a altura em centímetros. Desse modo, quando a altura é medida em polegadas, podemos obter um valor mais elevado para o numerador $\Sigma(x_i - \bar{x})(y_i - \bar{y})$ da Equação 3.10 e, portanto, uma covariância maior – quando, de fato, a relação não se altera. Uma medida da relação entre duas variáveis que não é afetada pelas unidades de medida para x e y é o **coeficiente de correlação**.

Figura 3.9 Interpretação da covariância da amostra

Coeficiente de Correlação

Para dados amostrais, o coeficiente de correlação momento-produto de Pearson é o seguinte:

COEFICIENTE DE CORRELAÇÃO MOMENTO-PRODUTO DE PEARSON: DADOS AMOSTRAIS

$$r_{xy} = \frac{s_{xy}}{s_x s_y} \tag{3.12}$$

em que

r_{xy} = coeficiente de correlação da amostra
s_{xy} = covariância da amostra
s_x = desvio padrão da amostra de x
s_y = desvio padrão da amostra de y

A Equação 3.12 mostra que o coeficiente de correlação momento-produto de Pearson para dados amostrais (comumente chamado *coeficiente de correlação da amostra*) é calculado dividindo-se a covariância da amostra pelo produto do desvio padrão da amostra de x pelo desvio padrão da amostra de y.

Agora, vamos calcular o coeficiente de correlação da amostra para a loja de equipamentos de som. Usando os dados da Tabela 3.8, podemos calcular os desvios padrão da amostra para as duas variáveis.

$$s_x = \sqrt{\frac{\Sigma(x_i - \bar{x})^2}{n - 1}} = \sqrt{\frac{20}{9}} = 1{,}49$$

$$s_y = \sqrt{\frac{\Sigma(y_i - \bar{y})^2}{n - 1}} = \sqrt{\frac{566}{9}} = 7{,}93$$

Assim, desde que $s_{xy} = 11$, o coeficiente de correlação da amostra é igual a

$$r_{xy} = \frac{s_{xy}}{s_x s_y} = \frac{11}{(1{,}49)(7{,}93)} = +{,}93$$

A fórmula para computar o coeficiente de correlação de uma população, denotado pela letra grega ρ_{xy} (pronuncia-se "rô"), é a seguinte:

COEFICIENTE DE CORRELAÇÃO MOMENTO-PRODUTO DE PEARSON:
DADOS POPULACIONAIS

$$\rho_{xy} = \frac{\sigma_{xy}}{\sigma_x \sigma_y} \tag{3.13}$$

em que

ρ_{xy} = coeficiente de correlação da população
σ_{xy} = covariância populacional
σ_x = desvio padrão da população para x
σ_y = desvio padrão da população para y

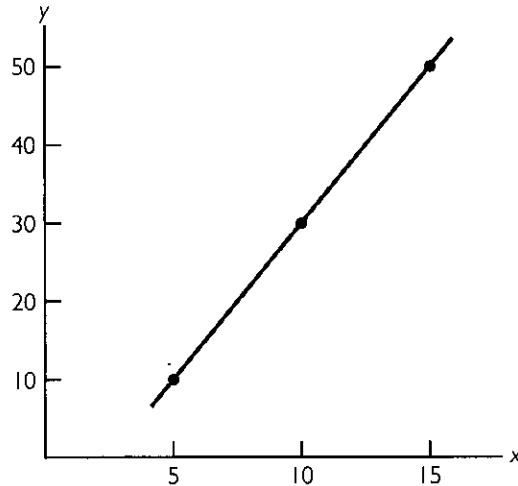
O coeficiente de correlação da amostra r_{xy} corresponde a uma estimativa do coeficiente de correlação da população ρ_{xy}.

O coeficiente de correlação da amostra r_{xy} fornece uma estimativa do coeficiente de correlação da população ρ_{xy}.

Interpretação do Coeficiente de Correlação

Primeiramente, vamos considerar um exemplo simples que ilustra o conceito de relação linear positiva perfeita. O diagrama de dispersão da Figura 3.10 descreve a relação entre x e y baseando-se nos seguintes dados amostrais:

x_i	y_i
5	10
10	30
15	50

Figura 3.10 Diagrama de dispersão descrevendo uma relação linear positiva perfeita

A linha reta traçada através de cada um dos três pontos indica uma relação linear perfeita entre x e y. Para aplicarmos a Equação 3.12 a fim de calcular a correlação da amostra, devemos primeiramente calcular s_{xy}, s_x e s_y. Alguns dos cálculos são mostrados na Tabela 3.9. Usando os resultados dessa tabela, encontramos

$$s_{xy} = \frac{\Sigma(x_i - \bar{x})(y_i - \bar{y})}{n - 1} = \frac{200}{2} = 100$$

$$s_x = \sqrt{\frac{\Sigma(x_i - \bar{x})^2}{n - 1}} = \sqrt{\frac{50}{2}} = 5$$

$$s_y = \sqrt{\frac{\Sigma(y_i - \bar{y})^2}{n - 1}} = \sqrt{\frac{800}{2}} = 20$$

$$r_{xy} = \frac{s_{xy}}{s_x s_y} = \frac{100}{5(20)} = 1$$

O coeficiente de correlação varia de −1 a +1. Valores que se aproximam de −1 ou +1 indicam uma relação linear forte. Quanto mais próxima a correlação estiver de zero, mais fraca será a relação.

Portanto, observamos que o valor do coeficiente de correlação da amostra é 1.

Em geral, pode-se demonstrar que, se todos os pontos de um conjunto de dados se situam em uma linha reta positivamente inclinada, o valor do coeficiente de correlação da amostra é +1; ou seja, um coeficiente de correlação da amostra igual a +1 corresponde a uma relação linear positiva perfeita entre x e y. Além disso, se os pontos do conjunto de dados se situam em uma linha reta que tem uma inclinação negativa, o valor do coeficiente de correlação da amostra é −1; ou seja, um coeficiente de correlação da amostra igual a −1 corresponde a uma relação linear negativa perfeita entre x e y.

Vamos supor agora que certo conjunto de dados indique uma relação linear positiva entre x e y, mas a relação não é perfeita. O valor de r_{xy} será menor que 1, indicando que os pontos no diagrama de dispersão não estão todos em uma linha reta. Uma vez que os pontos se afastam cada vez mais de uma relação linear positiva, o valor de r_{xy} torna-se cada vez menor. Um valor de r_{xy} igual a zero indica que não há nenhuma relação linear entre x e y, e os valores de r_{xy} próximos de zero indicam uma relação linear fraca.

Em relação aos dados envolvendo a loja de equipamentos de som, lembre-se de que r_{xy} = +0,93. Portanto, concluímos que ocorre uma relação linear positiva forte entre o número de comerciais e as vendas. Mais especificamente, um aumento no número de comerciais está associado a um aumento nas vendas.

Para encerrar, observamos que a correlação constitui uma medida de associação linear e não necessariamente de causação. Uma correlação elevada entre duas variáveis não significa que alterações havidas em uma variável provocarão alterações na outra variável. Por exemplo, podemos descobrir que a avaliação da qualidade e o preço típico das refeições em restaurantes estão positivamente correlacionados. Entretanto, simplesmente aumentar o preço em um restaurante não fará com que a avaliação da qualidade se eleve.

Tabela 3.9 Cálculos realizados para determinação do coeficiente de correlação da amostra

	x_i	y_i	$x_i - \bar{x}$	$(x_i - \bar{x})^2$	$y_i - \bar{y}$	$(y_i - \bar{y})^2$	$(x_i - \bar{x})(y_i - \bar{y})$
	5	10	−5	25	−20	400	100
	10	30	0	0	0	0	0
	15	50	5	25	20	400	100
Totais	30	90	0	50	0	800	200

$\bar{x} = 10 \quad \bar{y} = 30$

Exercícios

Métodos

AUTOTESTE

45. Cinco observações feitas de duas variáveis são apresentadas a seguir:

x_i	4	6	11	3	16
y_i	50	50	40	60	30

a. Desenvolva um diagrama de dispersão com x no eixo horizontal.
b. O que o diagrama de dispersão desenvolvido no item (a) indica a respeito da relação entre as duas variáveis?
c. Calcule e interprete a covariância da amostra.
d. Calcule e interprete o coeficiente de correlação da amostra.

46. Cinco observações feitas de duas variáveis são apresentadas a seguir:

x_i	6	11	15	21	27
y_i	6	9	6	17	12

a. Desenvolva um diagrama de dispersão desses dados.
b. O que o diagrama de dispersão indica a respeito da relação entre x e y?
c. Calcule e interprete a covariância da amostra.
d. Calcule e interprete o coeficiente de correlação da amostra.

Aplicações

47. A Nielsen Media Research publica duas medidas do público telespectador: uma *classificação* do programa de televisão, com a porcentagem dos lares que estão com os televisores ligados, e o *nível de audiência* do programa de televisão, contendo a porcentagem dos lares que assistem a determinado programa entre aqueles que estão com o televisor ligado. Os dados a seguir mostram classificação e os níveis de audiência referentes à transmissão dos jogos da Major League Baseball World Series ao longo de um período de nove anos (*Associated Press*, 27 de outubro de 2003):

Classificação	19	17	17	14	16	12	15	12	13
Nível de Audiência	32	28	29	24	26	20	24	20	22

a. Desenvolva um diagrama de dispersão com a classificação no eixo horizontal.
b. Qual é a relação entre a classificação e o nível de audiência? Explique.
c. Calcule e interprete a covariância da amostra.
d. Calcule o coeficiente de correlação da amostra. O que esse valor nos diz a respeito da relação entre a classificação e o nível de audiência?

48. Um estudo do departamento de transportes sobre a velocidade ao volante e a milhagem de automóveis de tamanho médio resultou nos seguintes dados:

Velocidade ao Volante	30	50	40	55	30	25	60	25	50	55
Milhagem	28	25	25	23	30	32	21	35	26	25

Calcule e interprete o coeficiente de correlação da amostra.

49. A revista *PC World* publicou avaliações de 15 computadores *notebook* (*PC World*, fevereiro de 2000). A pontuação do desempenho é uma medida de como o computador roda uma variedade de aplicativos comuns de negócios em comparação com uma máquina de referência. Por exemplo, um PC com um desempenho igual a 200 é duas vezes mais rápido que a máquina de referência. Foi utilizada uma escala de 100 pontos para fornecer uma avaliação global de cada *notebook* testado nesse estudo. Uma pontuação na casa dos 90 é excepcional, ao passo que uma pontuação na casa dos 70 é considerada boa. A Tabela 3.10 apresenta as pontuações de desempenho e as classificações gerais dos 15 *notebooks*.

Tabela 3.10 Pontuações de desempenho e classificações globais de 15 computadores *notebook*

ARQUIVO DA INTERNET
PCs

Notebook	Pontuação de Desempenho	Classificação Global
AMS Tech Roadster 15CTA380	115	67
Compaq Armada M700	191	78
Compaq Prosignia Notebook 150	153	79
Dell Inspiron 3700 C466GT	194	80
Dell Inspiron 7500 R500VT	236	84
Dell Latitude Cpi A366XT	184	76
Enpower ENP-313 Pro	184	77
Gateway Solo 9300LS	216	92
HP Pavilion Notebook PC	185	83
IBM ThinkPad I Series 1480	183	78
Micro Express NP7400	189	77
Micron TransPort NX PII-400	202	78
NEC Versa SX	192	78
Sceptre Soundx 5200	141	73
Sony VAIO PCG-F340	187	77

a. Calcule o coeficiente de correlação da amostra.
b. O que o coeficiente de correlação da amostra nos informa a respeito da relação entre a pontuação de desempenho e a classificação global?

50. A Média Industrial Dow Jones (*Dow Jones Industrial Average* – DJIA) e o Standard & Poor's (S&P) 500 Index são ambos utilizados como medidas do movimento global no mercado financeiro. A DJIA baseia-se no movimento de preços de 30 grandes empresas; o S&P 500 é um índice composto de 500 títulos financeiros. Alguns dizem que o S&P 500 é uma medida melhor do desempenho do mercado financeiro porque ele tem uma base mais ampla. Os preços de fechamento da DJIA e do S&P 500 correspondentes a dez semanas, com início em 11 de fevereiro de 2000, são mostrados a seguir (*Barron's*, 17 de abril de 2000).

Data	Dow Jones	S&P 500	Data	Dow Jones	S&P 500
11 de fevereiro	10.425	1.387	17 de março	10.595	1.464
18 de fevereiro	10.220	1.346	24 de março	11.113	1.527
25 de fevereiro	9.862	1.333	31 de março	10.922	1.499
3 de março	10.367	1.409	7 de abril	11.111	1.516
10 de março	9.929	1.395	14 de abril	10.306	1.357

ARQUIVO DA INTERNET
Dow S&P

a. Calcule o coeficiente de correlação da amostra desses dados.
b. Discuta a associação entre a DJIA e o S&P 500 Index.

51. As temperaturas máxima e mínima do dia (expressas em graus centígrados) de 12 cidades norte-americanas são apresentadas a seguir (Weather Channel, 25 de janeiro de 2004):

ARQUIVO DA INTERNET
Temperature

Cidade	Máxima	Mínima	Cidade	Máxima	Mínima
Albany	9	-8	Los Angeles	62	47
Boise	32	26	New Orleans	71	55
Cleveland	21	19	Portland	43	36
Denver	37	10	Providence	18	8
Des Moines	24	16	Raleigh	28	24
Detroit	20	17	Tulsa	55	38

a. Qual é a média de temperatura máxima diária da amostra?
b. Qual é a média de temperatura mínima diária da amostra?
c. Qual é a correlação entre as temperaturas máxima e mínima?

3.6 MÉDIA PONDERADA E O TRABALHO COM DADOS AGRUPADOS

Na Seção 3.1, apresentamos a média como uma das medidas mais importantes da posição central. A fórmula para encontrar a média de uma amostra com n observações é reformulada da seguinte maneira:

$$\bar{x} = \frac{\Sigma x_i}{n} = \frac{x_1 + x_2 + \cdots + x_n}{n} \qquad \textbf{(3.14)}$$

Nessa fórmula, cada x_i tem igual importância ou peso. Não obstante essa prática ser a mais comum, em alguns casos a média é calculada dando-se a cada observação um peso que reflita a sua importância. Uma média calculada dessa maneira é chamada **média ponderada**.

Média Ponderada

A média ponderada é calculada da seguinte maneira:

MÉDIA PONDERADA

$$\bar{x} = \frac{\Sigma w_i x_i}{\Sigma w_i} \qquad \textbf{(3.15)}$$

em que

x_i = o valor da observação i
w_i = o peso da observação i

Quando os dados são de uma amostra, a Equação 3.15 fornece a média ponderada da amostra. Quando os dados são de uma população, m substitui $\bar{x}$ e a equação fornece a média ponderada da população.

Como um exemplo da necessidade de se ter uma média ponderada, considere a seguinte amostra de cinco compras de determinada matéria-prima ao longo dos últimos três meses:

Compra	Custo por Quilo ($)	Quantidade em Quilos
1	3,00	1.200
2	3,40	500
3	2,80	2.750
4	2,90	1.000
5	3,25	800

Observe que o custo por quilo varia de US$ 2,80 a US$ 3,40 e que a quantidade comprada varia de 500 a 2.750 quilos. Suponha que um gerente tenha solicitado informações sobre o custo médio por quilo da matéria-prima. Uma vez que as quantidades encomendadas variam, precisamos usar a fórmula para a média ponderada. Os cinco valores de dados de custo por quilo são x_1 = 3,00, x_2 = 3,40, x_3 = 2,80, x_4 = 2,90 e x_5 = 3,25. A média ponderada do custo por quilo é encontrada ponderando-se cada custo por sua quantidade correspondente.

Para esse exemplo, os pesos são $w_1 = 1.200$, $w_2 = 500$, $w_3 = 2.750$, $w_4 = 1.000$ e $w_5 = 800$. Usando-se a Equação 3.15, a média ponderada é calculada da seguinte maneira:

$$\bar{x} = \frac{1.200(3{,}00) + 500(3{,}40) + 2.750(2{,}80) + 1.000(2{,}90) + 800(3{,}25)}{1.200 + 500 + 2.750 + 1.000 + 800}$$

$$= \frac{18.500}{6.250} = 2{,}96$$

Dessa forma, o cálculo da média ponderada mostra que o custo por quilo de matéria-prima é US$ 2,96. Note que o uso da Equação 3.14, em vez da fórmula da média ponderada, nos forneceria resultados enganosos. Nesse caso, a média dos cinco valores de custo por quilo é (3,00 + 3,40 + 2,80 + 2,90 + 3,25)/5 = 15,35/5 = US$ 3,07, a qual superestima o custo médio real por quilo comprado.

A escolha dos pesos para o cálculo de uma média ponderada em particular depende da aplicação. Um exemplo muito conhecido dos estudantes universitários norte-americanos é o cálculo da média escolar, a *grade point average* (GPA).[8] Nesse cálculo, os valores de dados geralmente usados são 4 para o grau A, 3 para o grau B, 2 para o grau C, 1 para o grau D e 0 para o grau F. Os pesos são o número horas-crédito conquistadas para cada grau. O Exercício 54 no fim desta seção apresenta um exemplo desse cálculo da média ponderada. Em outros cálculos da média ponderada, quantidades como libras-peso, dólares ou volume freqüentemente são utilizadas como pesos. De qualquer forma, quando as observações variam em termos de importância o analista deve escolher o peso que reflita melhor a importância de cada observação na determinação da média.

O cálculo da *grade point average* (GPA) é um bom exemplo do uso de uma média ponderada.

Dados Agrupados

Na maioria dos casos, as medidas de posição e variabilidade são calculadas usando-se os valores individuais dos dados. Às vezes, no entanto, os dados estão disponíveis somente na forma agrupada ou na forma de distribuição de freqüência. Na discussão a seguir, mostramos como a fórmula da média ponderada pode ser usada para se obter aproximações da média, da variância e do desvio padrão de **dados agrupados**.

Na Seção 2.2, apresentamos uma distribuição de freqüência do tempo em dias necessário para a conclusão das auditorias de fim de ano realizadas pela empresa de contabilidade Sanderson and Clifford. A distribuição de freqüência dos tempos para a conclusão das auditorias, baseada em uma amostra de 20 clientes, é indicada novamente na Tabela 3.11. Com base nessa distribuição de freqüência, qual é o tempo médio para conclusão das auditorias relativo à amostra?

Para calcular a média usando somente os dados agrupados, tratamos o ponto médio de cada classe como representativo dos itens da classe. Digamos que M_i denote o ponto médio da classe i e que f_i designe a freqüência da classe i. A fórmula da média ponderada (Equação 3.15) é então usada com os valores de dados denotados por M_i e os pesos pelas freqüências f_i. Nesse caso, o denominador da Equação 3.15 é a soma das freqüências, que é o tamanho n da amostra.

Tabela 3.11 Distribuição de freqüência dos tempos necessários para conclusão das auditorias

Tempo Necessário para a Conclusão das Auditorias (dias)	Freqüência
10–14	4
15–19	8
20–24	5
25–29	2
30–34	1
Total	20

Ou seja, $\sum f_i = n$. Desse modo, a equação da média da amostra para dados agrupados é a seguinte:

[8] NT: Média escolar nos Estados Unidos. Medida numérica do rendimento acadêmico baseada no cálculo do número de créditos e notas obtidas em todas as matérias até o presente. Baseia-se em uma escala de 0 a 4.

MÉDIA DA AMOSTRA PARA DADOS AGRUPADOS

$$\bar{x} = \frac{\Sigma f_i M_i}{n} \quad (3.16)$$

em que

M_i = o ponto médio da classe i
f_i = a freqüência da classe i
n = o tamanho da amostra

Com os pontos médios da classe, M_i, em uma posição intermediária entre os limites da classe, a primeira classe de 10–14 da Tabela 3.11 tem o ponto médio em (10 + 14)/2 = 12. Os cinco pontos médios da classe e o cálculo da média ponderada dos tempos para conclusão das auditorias estão resumidos na Tabela 3.12. Como se pode ver, a média de tempo para conclusão das auditorias da amostra são 19 dias.

Para calcular a variância de dados agrupados usamos uma versão ligeiramente modificada da fórmula para a variância apresentada na Equação 3.5. Nessa equação, os desvios dos dados em torno da média da amostra ao quadrado, $\bar{x}$, foram apresentados como $(x_i - \bar{x})^2$. Entretanto, com dados agrupados, os valores não são conhecidos. Nesse caso, tratamos o ponto médio da classe, M_i, como representativo dos valores x_i da classe correspondente. Desse modo, os desvios quadráticos em torno da média da amostra, $(x_i - \bar{x})^2$, são substituídos por $(M_i - \bar{x})^2$. Então, da mesma forma que agimos com os cálculos da média da amostra para dados agrupados, ponderamos cada valor pela freqüência de classe, f_i. A soma dos desvios quadráticos em torno da média de todos os dados é aproximada por $\Sigma f_i(M_i - \bar{x})^2$. O termo $n - 1$ em vez de n aparece no denominador a fim de transformar a variância da amostra na estimativa da variância da população. Assim, a fórmula apresentada a seguir é usada para se obter a variância da amostra de dados agrupados.

VARIÂNCIA DA AMOSTRA PARA DADOS AGRUPADOS

$$s^2 = \frac{\Sigma f_i(M_i - \bar{x})^2}{n - 1} \quad (3.17)$$

Tabela 3.12 Cálculo da média da amostra para dados agrupados do tempo necessário para conclusão das auditorias

Tempo Necessário para a Conclusão das Auditorias (dias)	Ponto Médio da Classe (M_i)	Freqüência (f_i)	f_iM_i
10–14	12	4	48
15–19	17	8	136
20–24	22	5	110
25–29	27	2	54
30–34	32	1	32
		20	380

Média da amostra $x = \frac{\Sigma f_i M_i}{n} = \frac{380}{20} = 19$ dias

Tabela 3.13 Cálculo da variância da amostra para dados agrupados do tempo necessário para a conclusão das auditorias (média da amostra $\bar{x} = 19$)

Tempo Necessário para a Conclusão das Auditorias (dias)	Ponto Médio da Classe (M_i)	Freqüência (f_i)	Desvio ($M_i - \bar{x}$)	Desvio Quadrático ($M_i - \bar{x}$)^2	$f_i(M_i - \bar{x})^2$
10–14	12	4	−7	49	196
15–19	17	8	−2	4	32
20–24	22	5	3	9	45
25–29	27	2	8	64	128
30–34	32	1	13	169	169
		20			570
					$\Sigma f_i(M_i - \bar{x})^2$

$$\text{Variância da amostra } s^2 = \frac{Sf_i(M_i - x)}{n - 1} = \frac{570}{19} = 30$$

O cálculo da variância da amostra dos tempos para conclusão das auditorias baseado nos dados agrupados da Tabela 3.11 é apresentado na Tabela 3.13. Como se pode notar, a variância da amostra é 30.

O desvio padrão de dados agrupados é simplesmente a raiz quadrada da variância dos dados agrupados. Em relação aos tempos para a conclusão das auditorias, o desvio padrão da amostra é $s = \sqrt{30} = 5{,}48$.

Antes de encerrarmos esta seção sobre o cálculo de medidas de posição e dispersão de dados agrupados, observamos que as Fórmulas 3.16 e 3.17 são para uma amostra. As medidas de sumário da população são computadas similarmente. As fórmulas de dados agrupados da média e da variância de uma população são apresentadas a seguir.

MÉDIA POPULACIONAL PARA DADOS AGRUPADOS

$$\mu = \frac{\Sigma f_i M_i}{N} \qquad \textbf{(3.18)}$$

VARIÂNCIA POPULACIONAL PARA DADOS AGRUPADOS

$$\sigma^2 = \frac{\Sigma f_i (M_i - \mu)^2}{N} \qquad \textbf{(3.19)}$$

NOTAS E COMENTÁRIOS

Ao calcularmos a estatística descritiva para dados agrupados, utilizamos os pontos médios da classe para aproximar os valores de dados de cada classe. Em conseqüência, a estatística descritiva para dados agrupados é uma aproximação da estatística descritiva que resultaria se usássemos os dados originais diretamente. Portanto, recomendamos calcular a estatística descritiva a partir dos dados originais em vez dos dados agrupados, sempre que isso for possível.

Exercícios

Métodos

52. Considere os seguintes dados e os pesos correspondentes:

x_i	Peso (w_i)
3,2	6
2,0	3
2,5	2
5,0	8

a. Calcule a média ponderada.
b. Calcule a média da amostra dos quatro valores de dados sem ponderação. Observe a diferença nos resultados apresentados pelos dois cálculos.

AUTOTESTE

53. Considere os dados amostrais da seguinte distribuição de freqüência:

Classe	Ponto Médio	Freqüência
3–7	5	4
8–12	10	7
13–17	15	9
18–22	20	5

a. Calcule a média da amostra.
b. Calcule a variância da amostra e o desvio padrão da amostra.

Aplicações

AUTOTESTE

54. A *grade point average* (GPA) dos estudantes universitários norte-americanos baseia-se no cálculo de uma média ponderada. Na maioria das escolas de ensino superior (*colleges*) são atribuídos os seguintes valores aos graus: A (4), B (3), C (2), D (1) e F (0). Depois de 60 horas-crédito de trabalho acadêmico, um estudante de uma universidade pública ganhou 9 horas-crédito para A, 15 horas-crédito para B, 33 horas-crédito para C e 3 horas-crédito para D.
   a. Calcule a GPA (*grade point average*) do estudante.
   b. Os estudantes das universidades públicas precisam manter uma *grade point average* de 2,5 para suas primeiras 60 horas-crédito de trabalho acadêmico a fim de serem admitidos na escola de administração. Esse estudante será admitido?

55. A *Bloomberg Personal Finance* (julho/agosto de 2001) incluiu as seguintes empresas em sua carteira de investimentos recomendada. Para uma carteira de US$ 25 mil, os valores em dólares que recomendavam alocar a cada ação são mostrados a seguir:

Empresa	Portfólio ($)	Estimativa da Taxa de Crescimento (%)	Retorno em Dividendos (%)
Citigroup	3.000	15	1,21
General Electric	5.500	14	1,48
Kimberly-Clark	4.200	12	1,72
Oracle	3.000	25	0,00
Pharmacia	3.000	20	0,96
SBC Communications	3.800	12	2,48
WorldCom	2.500	35	0,00

   a. Usando a quantia em dólares da carteira de investimentos como pesos, qual é a média ponderada da estimativa da taxa de crescimento da carteira de investimentos?
   b. Qual é a média ponderada do retorno em dividendos da carteira de investimentos?

56. Um posto de gasolina registrou a seguinte distribuição de freqüência para o número de galões de gasolina vendidos por carro em uma amostra de 680 carros.

Gasolina (galões)	Freqüência
0–4	74
5–9	192
10–14	280
15–19	105
20–24	23
25–29	6
Total	680

Calcule a média, a variância e o desvio padrão desses dados agrupados. Se o posto de gasolina espera atender a cerca de 120 carros em determinado dia, estime o número total de galões de gasolina que serão vendidos.

57. Uma pesquisa dos assinantes da revista *Fortune* fez a seguinte pergunta: "Quantas das quatro últimas edições você leu?" Suponha que a seguinte distribuição de freqüência resuma 500 respostas:

Número de Edições Lidas	Freqüência
0	15
1	10
2	40
3	85
4	350
Total	500

a. Qual é o número médio de edições lidas por um assinante da revista *Fortune*?
b. Qual é o desvio padrão do número de edições lidas?

Resumo

Neste capítulo, apresentamos diversos métodos de estatística descritiva que podem ser usados para sintetizar a posição, a variabilidade e a forma de uma distribuição de dados. Diferentemente dos procedimentos tabulares e gráficos introduzidos no Capítulo 2, as medidas inseridas neste capítulo sintetizam os dados em termos de valores numéricos. Quando os valores numéricos obtidos se referem a uma amostra, eles são chamados estatística da amostra. Quando os valores numéricos dizem respeito a uma população, eles são denominados parâmetros populacionais. Algumas das notações usadas para a estatística da amostra e para os parâmetros populacionais são:

	Estatística da Amostra	Parâmetro Populacional
Média	$\bar{x}$	μ
Variância	s^2	σ^2
Desvio padrão	s	σ
Covariância	s_{xy}	σ_{xy}
Correlação	r_{xy}	ρ_{xy}

Em inferência estatística, a estatística da amostra é chamada estimador por pontos do parâmetro populacional.

Como medidas da posição central, definimos a média, a mediana e a moda. Depois, utilizamos o conceito de percentil para descrever outras posições no conjunto de dados. Em seguida, apresentamos a amplitude, a amplitude interquartil, a variância, o desvio padrão e o coeficiente de variação como medidas da variabilidade ou dispersão. Nossa principal medida da forma de uma distribuição foi a assimetria. Valores negativos indicam uma distribuição de dados inclinada à esquerda. Valores positivos apontam uma distribuição de dados inclinada à direita. Logo após, descrevemos como a média e o desvio padrão poderiam ser usados, aplicando-se o teorema de Chebyshev e a regra empírica, para produzir informações mais específicas a respeito da distribuição de dados e para identificar os pontos fora da curva.

Na Seção 3.4, mostramos como desenvolver uma regra de cinco itens e um desenho esquemático (*box plot*) para fornecer informações simultâneas sobre a posição, variabilidade e forma da distribuição. Na Seção 3.5, introduzimos a covariância e o coeficiente de correlação como medidas da associação entre duas variáveis. Na seção final, mostramos como calcular uma média ponderada e como calcular uma média, a variância e o desvio padrão para dados agrupados.

A estatística descritiva que discutimos pode ser desenvolvida usando-se softwares estatísticos e planilhas eletrônicas. No Apêndice 3.1, mostraremos como desenvolver a maioria dos métodos de estatística descritiva apresentados neste capítulo, usando o Minitab. No Apêndice 3.2, demonstraremos o uso do Excel para o mesmo propósito.

Glossário

Estatística da amostra Valor numérico usado como medida resumida de uma amostra (por exemplo, a média da amostra, $\bar{x}$, a variância da amostra, s^2, e o desvio padrão da amostra, s).

Parâmetro populacional Valor numérico usado como medida resumida de uma população (por exemplo, a média populacional m, a variância de população s^2 e o desvio padrão s).

Estimador por pontos A estatística da amostra, por exemplo, $\bar{x}$, s^2 e s, quando usados para estimar o parâmetro populacional correspondente.

Média Medida de posição central que é calculada somando-se os valores de dados e dividindo-se o resultado pelo número de observações.

Mediana Medida de posição central fornecida pelo valor intermediário quando os dados são organizados em ordem crescente.

Moda Medida de posição, definida como o valor que ocorre com maior freqüência.

Percentil Valor tal que pelo menos p por cento das observações são menores ou iguais a esse valor e pelo menos $(100 - p)$ por cento das observações são maiores ou iguais a esse valor. O 50º percentil é a mediana.

Quartis O 25º, o 50º e o 75º percentis se denominam primeiro quartil, segundo quartil (mediana) e terceiro quartil, respectivamente. Os quartis podem ser usados para dividir um conjunto de dados em quatro partes, sendo cada parte com aproximadamente 25% dos dados.

Amplitude Medida de variabilidade, definida como o maior valor menos o menor valor.

Amplitude interquartil (AIQ) Medida de variabilidade, definida como a diferença entre o terceiro e o primeiro quartis.

Variância Medida de variabilidade baseada nos desvios dos valores de dados ao redor da média elevados ao quadrado.

Desvio padrão Medida de variabilidade calculada encontrando-se a raiz quadrada positiva da variância.

Coeficiente de variação Medida de variabilidade relativa calculada dividindo-se o desvio padrão pela média e multiplicando-se o resultado por 100.

Assimetria Medida da forma assumida por uma distribuição de dados. Dados inclinados à esquerda resultam em uma assimetria negativa; uma distribuição de dados simétrica resulta em uma simetria nula; e dados inclinados à direita resultam em uma simetria positiva.

Contagem-*z* Um valor encontrado dividindo-se o desvio ao redor da média $(x_i - \bar{x})$ pelo desvio padrão s. Uma contagem-*z* é chamada valor padronizado e denota o número de desvios padrão que x_i está afastado da média.

Teorema de Chebyshev Teorema que pode ser usado para se fazer afirmações acerca das propriedades dos valores de dados que devem estar contidos em um número específico de desvios padrão da média.

Regra empírica Regra que pode ser usada para calcular a porcentagem de valores de dados que devem estar dentro de um, dois e três desvios padrão da média para dados que exibem uma distribuição em forma de sino.

Ponto fora da curva Valor de dados incomumente pequeno ou incomumente grande.

Regra de cinco itens Técnica de análise exploratória de dados que usa cinco números para sintetizar os dados: o menor valor, o primeiro quartil, a mediana, o terceiro quartil e o maior valor.

Desenho esquemático (*box plot*) Sumário gráfico de dados que se baseia em uma regra de cinco itens.

Covariância Uma medida da associação linear entre duas variáveis. Valores positivos indicam uma relação positiva; valores negativos indicam uma relação negativa.

Coeficiente de correlação Medida de associação linear entre duas variáveis que assumem valores entre −1 e +1. Valores próximos de +1 indicam uma forte relação linear positiva; valores próximos de −1 indicam uma forte relação linear negativa; e valores próximos de zero indicam que não há nenhuma relação linear.

Média ponderada Média obtida atribuindo-se a cada observação um peso que reflete sua importância.

Dados agrupados Dados disponíveis em intervalos de classe quando sintetizados por uma distribuição de freqüência. Os valores individuais dos dados originais não estão disponíveis.

Fórmulas-Chave

Média da Amostra

$$\bar{x} = \frac{\Sigma x_i}{n} \tag{3.1}$$

Média Populacional

$$\mu = \frac{\Sigma x_i}{N} \tag{3.2}$$

Amplitude Interquartil

$$\text{AIQ} = Q_3 - Q_1 \tag{3.3}$$

Variância da População

$$\sigma^2 = \frac{\Sigma(x_i - \mu)^2}{N} \tag{3.4}$$

Variância da Amostra

$$s^2 = \frac{\Sigma(x_i - \bar{x})^2}{n - 1} \tag{3.5}$$

Desvio Padrão

$$\text{Desvio padrão da amostra} = s = \sqrt{s^2} \tag{3.6}$$

$$\text{Desvio padrão da população} = \sigma = \sqrt{\sigma^2} \tag{3.7}$$

Coeficiente de Variação

$$\left(\frac{\text{Desvio padrão}}{\text{Média}} \times 100\right)\% \tag{3.8}$$

Contagem-z

$$z_i = \frac{x_i - \bar{x}}{s} \tag{3.9}$$

Covariância da Amostra

$$s_{xy} = \frac{\Sigma(x_i - \bar{x})(y_i - \bar{y})}{n - 1} \tag{3.10}$$

Covariância Populacional

$$\sigma_{xy} = \frac{\Sigma(x_i - \mu_x)(y_i - \mu_y)}{N} \tag{3.11}$$

Coeficiente de Correlação Momento-Produto de Pearson: Dados Amostrais

$$r_{xy} = \frac{s_{xy}}{s_x s_y} \tag{3.12}$$

Coeficiente de Correlação Momento-Produto de Pearson: Dados Populacionais

$$\rho_{xy} = \frac{\sigma_{xy}}{\sigma_x \sigma_y} \tag{3.13}$$

Média Ponderada

$$\bar{x} = \frac{\Sigma w_i x_i}{\Sigma w_i} \tag{3.15}$$

Média da Amostra para Dados Agrupados

$$\bar{x} = \frac{\Sigma f_i M_i}{n} \tag{3.16}$$

Variância da Amostra para Dados Agrupados

$$s^2 = \frac{\Sigma f_i (M_i - \bar{x})^2}{n - 1} \tag{3.17}$$

Média Populacional para Dados Agrupados

$$\mu = \frac{\Sigma f_i M_i}{N} \tag{3.18}$$

Variância Populacional para Dados Agrupados

$$\sigma^2 = \frac{\Sigma f_i (M_i - \mu)^2}{N} \tag{3.19}$$

Exercícios Suplementares

ARQUIVO
DA INTERNET
Visa

58. De acordo com a Annual Consumer Spending Survey de 2003 a média mensal das taxas de cartão de crédito Visa do Bank of America foi de US$ 18,38 (*U.S. Airways Attaché Magazine*, dezembro de 2003). Uma amostra das taxas mensais de cartões de crédito apresenta os seguintes dados:

236	1.710	1.351	825	7.450
316	4.135	1.333	1.584	387
991	3.396	170	1.428	1.688

a. Calcule a média e a mediana.
b. Calcule o primeiro e o terceiro quartis.
c. Calcule a amplitude e a amplitude interquartil.
d. Calcule a variância e o desvio padrão.
e. A medida de assimetria desses dados é 2,12. Comente a forma dessa distribuição. Ela é a forma que você esperaria? Por quê? Por que não?
f. Os dados contêm pontos fora da curva?

59. A remuneração total anual dos membros da diretoria de uma das 100 maiores empresas de capital aberto do país se baseia parcialmente nos *cash retainers*,[9] um pagamento anual por participarem do quadro de diretores. Além dos *cash retainers*, os membros da diretoria recebem uma *stock retainer*, uma subvenção em ações, uma *stock option*[10] e honorários por participarem das reuniões da diretoria. A remuneração total pode facilmente ultrapassar os US$ 100 mil, até mesmo quando se tem *cash retainers* baixos, por exemplo, US$ 15 mil. Os dados a seguir apresentam o *cash retainer* (em US$ 1.000) correspondente a uma amostra de 20 das maiores empresas de capital aberto do país (*USA Today*, 17 de abril de 2000).

ARQUIVO
DA INTERNET
Retainer

Empresa	*Cash Retainer*
American Express	64
Bank of America	36
Boeing	26
Chevron	35
Dell Computer	40
DuPont	35
ExxonMobil	40
Ford Motor	30
General Motors	60
International Paper	36
Kroger	28
Lucent Technologies	50
Motorola	20
Procter & Gamble	55
Raytheon	40
Sears Roebuck	30
Texaco	15
United Parcel Service	55
Wal-Mart Stores	25
Xerox	40

Calcule a seguinte estatística descritiva:
a. A média, a mediana e a moda.
b. O primeiro e o terceiro quartis.
c. A amplitude e a amplitude interquartis.
d. A variância e o desvio padrão.
e. O coeficiente de variação.

[9] NT: Tipo de adiantamento em dinheiro para os participantes da diretoria.
[10] NT: Programa que permite aos empregados comprarem ações da empresa a preço e lucro fixos quando seu desempenho no mercado eleva o valor de suas ações.

60. O retorno em dividendos são os dividendos anuais que uma empresa paga, divididos pelo preço de mercado atual por ação, expressos na forma de porcentagem. Uma amostra de dez grandes empresas produziu os seguintes dados sobre o retorno em dividendos (*The Wall Street Journal*, 16 de janeiro de 2004).

Empresa	Retorno em Dividendos (%)	Empresa	Retorno em Dividendos (%)
Altria Group	5,0	General Motors	3,7
American Express	0,8	JPMorgan Chase	3,5
Caterpillar	1,8	McDonald's	1,6
Eastman Kodak	1,9	United Technology	1,5
ExxonMobil	2,5	Wal-Mart Stores	0,7

a. Qual é a média e a média dos retornos em dividendos?
b. Qual é a variância e o desvio padrão?
c. Qual empresa proporciona o maior retorno em dividendos?
d. Qual é a contagem-z do McDonald's? Interprete essa contagem-z.
e. Qual é a contagem-z da General Motors? Interprete essa contagem-z.
f. Com base nas contagens-z, os dados contêm algum ponto fora da curva?

61. De acordo com a Forrester Research Inc., aproximadamente 19% dos usuários da internet divertem-se com jogos on-line. Os dados a seguir mostram o número de usuários exclusivos (em milhares) de dez sites de jogos no mês de março (*The Wall Street Journal*, 17 de abril de 2000).

Site	Usuários Exclusivos
aolgames.com	9.416
extremelotto.com	3.955
freelotto.com	12.901
gamesville.com	4.844
iwin.com	7.410
prizecentral.com	4.899
shockwave.com	5.582
speedyclick.com	6.628
uproar.com	8.821
webstakes.com	7.499

Usando esses dados, calcule a média, a mediana, a variância e o desvio padrão.

62. A renda familiar típica de uma amostra de 20 cidades é apresentada a seguir (*Places Rated Almanac*, 2000). Os dados estão expressos em milhares de dólares:

ARQUIVO DA INTERNET
Income

Cidade	Renda
Akron, OH	74,1
Atlanta, GA	82,4
Birmingham, AL	71,2
Bismark, ND	62,8
Cleveland, OH	79,2
Columbia, SC	66,8
Danbury, CT	132,3
Denver, CO	82,6
Detroit, MI	85,3
Fort Lauderdale, FL	75,8
Hartford, CT	89,1
Lancaster, PA	75,2
Madison, WI	78,8
Naples, FL	100,0
Nashville, TN	77,3
Philadelphia, PA	87,0
Savannah, GA	67,8
Toledo, OH	71,2
Trenton, NJ	106,4
Washington, DC	97,4

a. Calcule a média e o desvio padrão dos dados da amostra.
b. Usando a média e o desvio padrão calculados no item (a) como estimativas da média e do desvio padrão da renda familiar da população de todas as cidades, use o teorema de Chebyshev para deter-

minar a amplitude dentro da qual 75% das rendas familiares da população de todas as cidades devem se situar.

c. Suponha que a distribuição da renda familiar tenha a forma de sino. Usando a média e o desvio padrão calculados no item (a) como estimativas da média e do desvio padrão da renda familiar da população de todas as cidades, use a regra empírica para determinar a amplitude dentro da qual 95% das rendas familiares da população de todas as cidades devem se situar. Compare sua resposta com o valor encontrado no item (b).

63. O transporte público e o automóvel são dois métodos que os trabalhadores podem usar para chegar ao trabalho diariamente. As amostras de tempo registradas para cada método são apresentadas a seguir. Os tempos estão expressos em minutos:

Transporte Público:	28	29	32	37	33	25	29	32	41	34
Automóvel:	29	31	33	32	34	30	31	32	35	33

a. Calcule a média de tempo da amostra para se chegar ao trabalho utilizando cada um dos meios de transporte.

b. Calcule o desvio padrão da amostra de cada meio de transporte.

c. Tendo como base os resultados que você obteve nos itens (a) e (b), qual meio de transporte deveria ser preferível? Explique.

d. Desenvolva um desenho esquemático (*box plot*) correspondente a cada meio de transporte. Uma comparação dos desenhos esquemáticos sustenta suas conclusões para o item (c)?

64. A renda familiar típica e o preço típico das casas em uma amostra de 20 cidades são os seguintes (*Places Rated Almanac*, 2000). Os dados estão expressos em milhares de dólares.

ARQUIVO DA INTERNET
Cities

Cidade	Renda	Preços das Casas
Bismark, ND	62,8	92,8
Columbia, SC	66,8	116,7
Savannah, GA	67,8	108,1
Birmingham, AL	71,2	130,9
Toledo, OH	71,2	101,1
Akron, OH	74,1	114,9
Lancaster, PA	75,2	125,9
Fort Lauderdale, FL	75,8	145,3
Nashville, TN	77,3	125,9
Madison, WI	78,8	145,2
Cleveland, OH	79,2	135,8
Atlanta, GA	82,4	126,9
Denver, CO	82,6	161,9
Detroit, MI	85,3	145,0
Philadelphia, PA	87,0	151,5
Hartford, CT	89,1	162,1
Washington, DC	97,4	191,9
Naples, FL	100,0	173,6
Trenton, NJ	106,4	168,1
Danbury, CT	132,3	234,1

a. Qual é o valor da covariância da amostra? Ela indica uma relação linear positiva ou negativa?

b. Qual é o coeficiente de correlação da amostra?

65. Os dados a seguir apresentam os gastos com a mídia (milhões de dólares) e as remessas em milhões de barris referentes a dez grandes marcas de cerveja.

ARQUIVO DA INTERNET
Beer

Marca	Gastos com a Mídia (milhões de dólares)	Remessa em Milhões de Barris
Budweiser	120,0	36,3
Bud Light	68,7	20,7
Miller Lite	100,1	15,9
Coors Light	76,6	13,2
Busch	8,7	8,1
Natural Light	0,1	7,1
Miller Genuine Draft	21,5	5,6
Miller High Life	1,4	4,4
Busch Lite	5,3	4,3
Milwaukee's Best	1,7	4,3

a. Qual é a covariância da amostra? Ela indica uma relação positiva ou negativa?
b. Qual é o coeficiente de correlação da amostra?

66. A *Road & Track* publicou a seguinte amostra de avaliações da vida útil e da capacidade de carga de pneus de automóvel:

Avaliação do Pneu	Capacidade de Carga
75	853
82	1.047
85	1.135
87	1.201
88	1.235
91	1.356
92	1.389
93	1.433
105	2.039

a. Desenvolva um diagrama de dispersão dos dados, colocando a classificação dos pneus no eixo x.
b. Qual é o coeficiente de correlação da amostra, e o que ele lhe informa sobre a relação entre a avaliação do pneu e a capacidade de carga?

67. Os dados seguintes mostram a rentabilidade de ações de primeira linha em um *trailing*[11] de 52 semanas e os valores nominais registrados por dez empresas (*The Wall Street Journal*, 13 de março de 2000).

Empresa	Valor Nominal	Rentabilidade
Am Elec	25,21	2,69
Columbia En	23,20	3,01
Con Ed	25,19	3,13
Duke Energy	20,17	2,25
Edison Int'l	13,55	1,79
Enron Cp.	7,44	1,27
Peco	13,61	3,15
Pub Sv Ent	21,86	3,29
Southn Co.	8,77	1,86
Unicom	23,22	2,74

a. Desenvolva um diagrama de dispersão dos dados, representando o valor nominal no eixo x.
b. Qual é o coeficiente de correlação da amostra, e o que ela lhe informa a respeito da relação entre a rentabilidade por ação e o valor nominal?

68. Uma técnica de previsão denominada média móvel usa a média, ou ponto médio, dos n períodos mais recentes para prever o valor seguinte dos dados de uma série temporal. Com uma média móvel de três períodos, os três períodos de dados mais recentes são utilizados no cálculo da previsão. Considere um produto com a seguinte demanda para os três primeiros meses do ano atual: janeiro (800 unidades), fevereiro (750 unidades) e março (900 unidades).

a. Qual é a previsão em termos de média móvel de três meses para abril?
b. Uma variação dessa técnica de previsão denomina-se média móvel ponderada. A ponderação possibilita que se atribua mais peso ou mais importância aos dados mais recentes da série temporal no cálculo da previsão. Por exemplo, uma média móvel ponderada de três meses poderia dar um peso 3 a dados de um mês atrás, peso 2 a dados de dois meses atrás e peso 1 a dados de três meses atrás. Use os dados apresentados para fornecer uma previsão em termos de média móvel ponderada de três meses para abril.

69. Os prazos em dias para a data de vencimento de uma amostra de cinco fundos de investimento são apresentados como segue. As quantias em dólares investidas nos fundos são fornecidas. Use a média ponderada para determinar o número médio de dias até a data de vencimento para os dólares investidos nesses cinco fundos de investimento.

[11] NT: Técnica utilizada para mover o preço de fechamento para um ponto próximo dos preços negociados à medida que estes seguem para a direção desejada. O objetivo é cortar perdas.

Prazo em Dias para a Data de Vencimento	Valores em Dólares (milhões)
20	20
12	30
7	10
5	15
6	10

70. Os automóveis que trafegam em uma rodovia com limite de velocidade fixado em 55 milhas por hora (88,51 km/h) têm a velocidade mostrada por um sistema de radares da polícia estadual. Uma distribuição da freqüência das velocidades é apresentada a seguir:

Velocidade (milhas por hora)		Freqüência
45–49		10
50–54		40
55–59		150
60–64		175
65–69		75
70–74		15
75–79		10
	Total	475

a. Qual é a velocidade média dos automóveis que trafegam nessa rodovia?
b. Calcule a variância e o desvio padrão.

Estudo de Caso 1 – Pelican Stores

A Pelican Stores, uma cadeia de lojas de vestuário feminino que opera em todos os Estados Unidos, realizou recentemente uma promoção na qual cupons de desconto eram enviados a clientes das lojas associadas. Dados coletados de uma amostra de 100 transações com cartão de crédito na loja durante um dia em novembro de 2002 se encontram no arquivo intitulado Pelican. A Tabela 3.14 apresenta uma parte do conjunto de dados. Um valor não-igual a zero para a variável desconto indica que a cliente trouxe os cupons promocionais e os usou. Para alguns clientes, o valor do desconto é maior que o valor das vendas (veja o Cliente 4). O valor das vendas é líquido, sem descontos ou trocos.

ARQUIVO DA INTERNET
Pelican

Tabela 3.14 Dados de uma amostra de 100 compras com cartões de crédito nas lojas Pelican

Cliente	Método de Pagamento	Artigos	Valor do Desconto	Vendas	Sexo	Estado Civil	Idade
1	Discover	1	0,00	39,50	Masculino	Casado	32
2	Proprietary Card	1	25,60	102,40	Feminino	Casada	36
3	Proprietary Card	1	0,00	22,50	Feminino	Casada	32
4	Proprietary Card	5	121,10	100,40	Feminino	Casada	28
5	Mastercard	2	0,00	54,00	Feminino	Casada	34
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.
96	Mastercard	1	0,00	39,50	Feminino	Casada	44
97	Proprietary Card	9	82,75	253,00	Feminino	Casada	30
98	Proprietary Card	10	18,00	287,59	Feminino	Casada	52
99	Proprietary Card	2	31,40	47,60	Feminino	Casada	30
100	Proprietary Card	1	11,06	28,44	Feminino	Casada	44

A administração das lojas Pelican quer usar essa amostra para conhecer sua base de clientes e para avaliar a promoção envolvendo cupons de desconto.

Relatório Administrativo

Use os métodos de estatística descritiva apresentados neste capítulo para sintetizar os dados e comente suas descobertas. No mínimo, seu relatório deve incluir o seguinte:

1. A estatística descritiva das vendas e a estatística descritiva das vendas de acordo com várias classificações de clientes.
2. A estatística descritiva da relação entre o valor do desconto e as vendas para os clientes que responderam à promoção.
3. A estatística descritiva da relação entre idade e vendas.

Comente quaisquer resultados que pareçam interessantes e de valor potencial para a administração.

Estudo de Caso 2 – National Health Care Association

A National Health Care Association está preocupada com a escassez de enfermeiras que o setor de enfermagem projeta para o futuro. Para saber qual é o grau atual de satisfação no trabalho entre as profissionais, a associação patrocinou um estudo das enfermeiras de hospital em todo o território nacional dos Estados Unidos. Como parte desse estudo, 50 enfermeiras de uma amostra indicaram seus níveis de satisfação com o trabalho, com seus salários e com suas oportunidades de promoção. Cada um dos três aspectos de satisfação foi medido em uma escala de 0 a 100, com os valores mais altos indicando níveis mais elevados de satisfação. Os dados coletados também mostraram os tipos de hospital que empregam as enfermeiras. Os tipos de hospital eram particular, *Veterans Administration* (VA)[12] e universitário. Uma parte dos dados se encontra na Tabela 3.15. O conjunto de dados completo pode ser encontrado no site www.thomsonlearning.com.br/estatapl.htm, no arquivo intitulado Health.

Tabela 3.15 Dados do nível de satisfação no trabalho de uma amostra de 50 enfermeiras

Enfermeira	Hospital	Trabalho	Remuneração	Promoção
1	Particular	74	47	63
2	*Veterans Administration* (VA)	72	76	37
3	Universitário	75	53	92
4	Particular	89	66	62
5	Universitário	69	47	16
6	Particular	85	56	64
7	Universitário	89	80	64
8	Particular	88	36	47
9	Universitário	88	55	52
10	Particular	84	42	66
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.
45	Universitário	79	59	41
46	Universitário	84	53	63
47	Universitário	87	66	49
48	*Veterans Administration* (VA)	84	74	37
49	*Veterans Administration* (VA)	95	66	52
50	Particular	72	57	40

ARQUIVO
DA INTERNET
Health

Relatório Administrativo

Use métodos de estatística descritiva para sintetizar os dados. Apresente sumários que sejam eficientes em termos de comunicar os resultados a outras pessoas. Discuta suas descobertas. Especificamente, comente as seguintes questões:

[12] NT: Órgão federal consolidado que administra todas as leis que regem os benefícios para veteranos das Forças Armadas.

1. Com base no conjunto de dados inteiro e nas três variáveis de satisfação no trabalho, qual aspecto do trabalho é o mais satisfatório para as enfermeiras? Qual parece ser o menos satisfatório? Em quais áreas, se for o caso, você acha que se devem fazer melhorias? Discuta o assunto.
2. Com base nas medidas descritivas de variabilidade, qual medida de satisfação no trabalho parece gerar a maior diferença de opinião entre as enfermeiras? Explique.
3. O que se pode aprender em relação aos tipos de hospital? Um tipo de hospital em particular parece apresentar melhores níveis de satisfação no trabalho que os outros? Os resultados que você obteve sugerem alguma recomendação para que se possa conhecer e melhorar a satisfação no trabalho? Discuta o assunto.
4. Qual estatística descritiva e *insights* adicionais você é capaz de usar para conhecer e possivelmente melhorar a satisfação no trabalho?

Estudo de Caso 3 – Escolas de Administração da Região Ásia-Pacífico

ARQUIVO DA INTERNET
Asian

A busca de um diploma de nível superior em administração agora é internacional. Uma pesquisa revela que um número cada vez maior de asiáticos optam pelo caminho da graduação em MBA – *Master of Business Administration* – para chegar ao sucesso corporativo (*Ásia, Inc.*, setembro de 1997). O número de candidatos em cursos de MBA em escolas da região Ásia-Pacífico continua a crescer cerca de 30% ao ano. Em 1997, as 74 escolas de administração da região Ásia-Pacífico registraram um recorde de 170 mil candidatos aos 11 mil diplomas de MBA de tempo integral (*full-time*) que seriam concedidos em 1999. Uma das razões principais para o crescimento da demanda é que um MBA pode aumentar substancialmente o poder remunerativo.

Em toda a região, milhares de asiáticos demonstram uma crescente disposição para interromper temporariamente suas carreiras e despender dois anos em busca de uma qualificação teórica em administração. Os cursos ministrados nessas escolas são notoriamente árduos e incluem economia, operações bancárias, marketing, ciências comportamentais, relações no trabalho, tomada de decisões, pensamento estratégico, direito comercial e outros. A *Ásia, Inc.*, forneceu o conjunto de dados da Tabela 3.16, a qual apresenta algumas das características das principais escolas de administração da região Ásia-Pacífico.

Relatório Administrativo

Use os métodos de estatística descritiva para sintetizar os dados da Tabela 3.16. Discuta suas descobertas.

1. Inclua um sumário correspondente a cada variável do conjunto de dados. Faça comentários e interpretações baseadas nos máximos e mínimos, bem como nas médias e proporções apropriadas. Quais novos *insights* essas estatísticas descritivas ofereceriam em relação às escolas de administração da região Ásia-Pacífico?
2. Sintetize os dados para comparar o seguinte:
   a. Quaisquer diferenças entre os custos de instrução no local e no exterior.
   b. Quaisquer diferenças entre a média dos salários iniciais das escolas que exigem experiência profissional e as que não exigem.
   c. Quaisquer diferenças entre os salários iniciais das escolas que exigem exames de inglês e as que não exigem.
3. Os salários iniciais parecem estar relacionados aos custos de instrução?

Apresente quaisquer sumários gráficos e numéricos adicionais que sejam benéficos em termos de comunicar os dados da Tabela 3.16 a outras pessoas.

Tabela 3.16 Dados de 25 escolas de administração da região Ásia-Pacífico

Escola de Administração	Matrícula em Curso de Tempo Integral	Número de Estudantes por Docente	Custo de Instrução Local	Custo de Instrução no Exterior	Idade	Porcentagem de Estrangeiros	GMAT[13]	Exame de Inglês	Experiência Profissional	Salário Inicial (US$)
Melbourne Business School	200	5	24.420	29.600	28	47	Sim	Não	Sim	71.400
University of New South Wales (Sydney)	228	4	19.993	32.582	29	28	Sim	Não	Sim	65.200
Indian Institute of Management (Ahmedabad)	392	5	4.300	4.300	22	0	Não	Não	Não	7.100
Chinese University of Hong Kong	90	5	11.140	11.140	29	10	Sim	Não	Não	31.000
International University of Japan (Niigata)	126	4	33.060	33.060	28	60	Sim	Sim	Não	87.000
Asian Institute of Management (Manila)	389	5	7.562	9.000	25	50	Sim	Não	Sim	22.800
Indian Institute of Management (Bangalore)	380	5	3.935	16.000	23	1	Sim	Não	Não	7.500
National University of Singapore	147	6	6.146	7.170	29	51	Sim	Sim	Sim	43.300
Indian Institute of Management (Calcutta)	463	8	2.880	16.000	23	0	Não	Não	Não	7.400
Australian National University (Canberra)	42	2	20.300	20.300	30	80	Sim	Sim	Sim	46.600
Nanyang Technological University (Singapore)	50	5	8.500	8.500	32	20	Sim	Não	Sim	49.300
University of Queensland (Brisbane)	138	17	16.000	22.800	32	26	Não	Não	Sim	49.600
Hong Kong University of Science and Technology	60	2	11.513	11.513	26	37	Sim	Não	Sim	34.000
Macquarie Graduate School of Management (Sydney)	12	8	17.172	19.778	34	27	Não	Não	Sim	60.100
Chulalongkorn University (Bangkok)	200	7	17.355	17.355	25	6	Sim	Não	Sim	17.600
Monash Mt. Eliza Business School (Melbourne)	350	13	16.200	22.500	30	30	Sim	Sim	Sim	52.500
Asian Institute of Management (Bangkok)	300	10	18.200	18.200	29	90	Não	Sim	Sim	25.000
University of Adelaide	20	19	16.426	23.100	30	10	Não	Não	Sim	66.000
Massey University (Palmerston North, New Zealand)	30	15	13.106	21.625	37	35	Não	Sim	Sim	41.400
Royal Melbourne Institute of Technology Business Graduate School	30	7	13.880	17.765	32	30	Não	Sim	Sim	48.900
Jamnalal Bajaj Institute of Management Studies (Bombay)	240	9	1.000	1.000	24	0	Não	Não	Sim	7.000
Curtin Institute of Technology (Perth)	98	15	9.475	19.097	29	43	Sim	Não	Sim	55.000
Lahore University of Management Sciences	70	14	11.250	26.300	23	2.5	Não	Não	Não	7.500
Universiti Sains Malaysia (Penang)	30	5	2.260	2.260	32	15	Não	Sim	Sim	16.000
De La Salle University (Manila)	44	17	3.300	3.600	28	3.5	Sim	Não	Sim	13.100

[13] NT: *Graduate Management Admission Test.*

Apêndice 3.1 – Estatística Descritiva com o Minitab

Neste Apêndice, descrevemos como usar o Minitab para desenvolver estatísticas descritivas. A Tabela 3.1 relacionou os salários iniciais de 12 diplomados da escola de administração. O painel A da Figura 3.11 apresenta a estatística descritiva obtida usando-se o Minitab para sintetizar esses dados. As definições dos cabeçalhos do Painel A são as seguintes:

N	Número de valores de dados
N*	Número de dados que faltam
Média	Média
EP da Média	Erro padrão da média
StDev	Desvio padrão
Mínimo	Valor mínimo de dados
Q1	Primeiro quartil
Mediana	Mediana
Q3	Terceiro quartil
Máximo	Valor máximo de dados

O rótulo "EP da Média" (na Tabela 3.16) refere-se ao *erro padrão da média*. Ele é calculado dividindo-se o desvio padrão pela raiz quadrada de *N*. A interpretação e o uso dessa medida serão discutidos no Capítulo 7, quando introduziremos o tema da amostragem e das distribuições de amostragem.

Não obstante as medidas numéricas de amplitude, amplitude interquartil, variância e coeficiente de variação não aparecerem na saída do Minitab, esses valores podem ser facilmente calculados a partir dos resultados contidos na Figura 3.11 da seguinte maneira:

$$\begin{aligned} \text{Amplitude} &= \text{Máximo} - \text{Mínimo} \\ \text{AIQ} &= Q_3 - Q_1 \\ \text{Variância} &= (\text{StDev})^2 \\ \text{Coeficiente de Variação} &= (\text{StDev/Média}) \times 100 \end{aligned}$$

Finalmente, observe que os quartis $Q_1 = 2.857{,}5$ e $Q_3 = 3.025$ do Minitab são ligeiramente diferentes dos quartis $Q_1 = 2.865$ e $Q_3 = 3.000$ calculados na Seção 3.1. As diferentes convenções* usadas para identificar os quartis explicam essa variação. Portanto, os valores de Q_1 e Q_3 fornecidos por uma convenção podem não ser idênticos aos valores de Q_1 e Q_3 fornecidos por outra convenção. Entretanto, quaisquer diferenças tendem a ser desprezíveis, e os resultados apresentados não induzirão os usuários a erro ao fazerem as interpretações habituais associadas aos quartis.

ARQUIVO DA INTERNET
Salary

Vejamos agora como as estatísticas da Figura 3.11 são geradas. Os dados de salários iniciais estão na coluna C2 de uma planilha do Minitab. As etapas a seguir podem ser usadas para gerar a estatística descritiva.

Etapa 1. Selecione o menu **Stat**
Etapa 2. Escolha **Basic Statistics**
Etapa 3. Escolha **Display Descriptive Statistics**
Etapa 4. Quando a caixa de diálogo Display Descriptive Statistics aparecer:
Digite C2 na caixa **Variables**
Dê um clique em **OK**

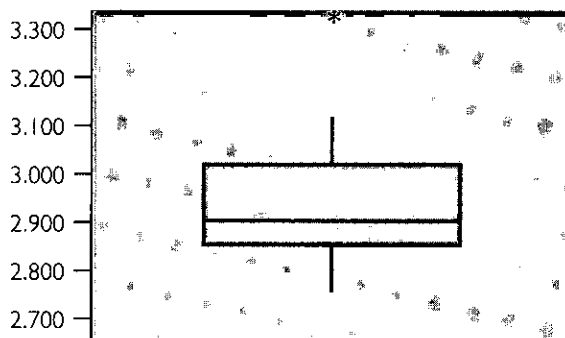
* Com as *n* observações organizadas em ordem crescente (do menor para o maior valor), o Minitab usa as posições fornecidas por $(n + 1)/4$ e $(3n + 1)/4$ para localizar Q_1 e Q_3, respectivamente. Quando uma posição é fracionária, o Minitab faz a interpolação entre os dois valores de dados dispostos em ordem adjacente para determinar o quartil correspondente.

Figura 3.11 Estatística descritiva e desenho esquemático produzidos pelo Minitab

Painel A: Estatística Descritiva

N	N*	Média	EP da Média	Desvio Padrão
12	0	2.940,0	47,8	165,7

Mínimo	Q1	Mediana	Q3	Máximo
2.710,0	2.857,5	2.905,0	3.025,0	3.325,0

Painel B: Desenho Esquemático (*Box plot*)

O Painel B da Figura 3.11 é um desenho esquemático produzido pelo Minitab. O retângulo traçado do primeiro ao terceiro quartis contém os 50% intermediários dos dados. A linha contida no retângulo assinala a mediana. O asterisco indica um ponto fora da curva em 3.325.

As etapas a seguir geram o desenho esquemático apresentado no Painel B da Figura 3.11.

Etapa 1. Selecione o menu **Graph**
Etapa 2. Escolha **Boxplot**
Etapa 3. Selecione **Simple** e dê um clique em **OK**
Etapa 4. Quando a caixa de diálogo Boxplot-One Y, Simple aparecer:
Digite C2 na caixa **Graph variables**
Dê um clique em **OK**

A medida de assimetria também não aparece como parte dos dados de saída (*output*) de estatística descritiva padrão do Minitab. Entretanto, podemos incluí-la na tela de estatística descritiva seguindo essas etapas:

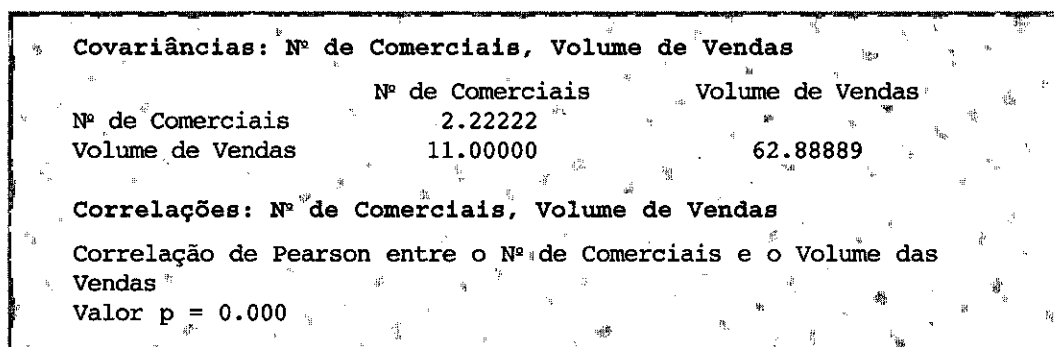
Etapa 1. Selecione o menu **Stat**
Etapa 2. Escolha **Basic Statistics**
Etapa 3. Escolha **Display Descriptive Statistics**
Etapa 4. Quando a caixa de diálogo Display Descriptive Statistics aparecer:
Dê um clique em **Statistics**
Selecione **Skewness**
Dê um clique em **OK**
Dê um clique em **OK**

A medida de assimetria 1,09 aparecerá então em sua planilha.

ARQUIVO DA INTERNET
Stereo

A Figura 3.12 mostra os dados de saída (*output*) de covariância e correlação que o Minitab produziu para os dados referentes à loja de equipamentos de som da Tabela 3.7. Na parte da figura relativa à covariância, nº de comerciais indica o número de comerciais de televisão de fins de semana, e volume de Vendas indica as vendas durante a semana seguinte. O valor 11 indicado na coluna nº de comerciais e na linha Volume de Vendas é a covariância da amostra, de acordo com o que foi calculado na Seção 3.5. O valor 2.22222 na coluna nº de comerciais e na linha nº de comerciais é a variância da amostra do número de comerciais, e o valor 62.88889 na coluna volume de vendas e na linha volume de vendas é a variância da amostra correspondente às vendas. O coeficiente de correlação da amostra, 0,930, é mostrado na parte correspondente à correlação nos dados de saída. *Nota*: A interpretação do valor $p = 0,000$ será discutida no Capítulo 9.

Figura 3.12 Covariância e correlação entre o número de comerciais e de vendas produzidas pelo Minitab

```
Covariâncias: Nº de Comerciais, Volume de Vendas

                    Nº de Comerciais    Volume de Vendas
Nº de Comerciais         2.22222
Volume de Vendas        11.00000            62.88889

Correlações: Nº de Comerciais, Volume de Vendas

Correlação de Pearson entre o Nº de Comerciais e o Volume das
Vendas
Valor p = 0.000
```

Vamos descrever, agora, como obter a informação da Figura 3.12. Introduzimos os dados referentes ao número de comerciais na coluna C2 e os dados referentes ao volume de vendas na coluna C3 de uma planilha Minitab. As etapas necessárias para gerar os dados de saída de covariância que foram apresentados nas três primeiras linhas da Figura 3.12 são mostradas a seguir:

Etapa 1. Selecione o menu **Stat**
Etapa 2. Escolha **Basic Statistics**
Etapa 3. Escolha **Covariance**
Etapa 4. Quando a caixa de diálogo Covariance aparecer:
Digite C2 C3 na caixa **Variables**
Dê um clique em **OK**

Para se obter os dados de saída (*output*) de correlação apresentados na Figura 3.12, somente uma mudança é necessária nas etapas destinadas à obtenção da covariância. Na etapa 3, a opção **Correlation** é selecionada.

Apêndice 3.2 – Estatística Descritiva com o Excel

O Excel pode ser usado para gerar a estatística descritiva discutida neste capítulo. Mostramos como o Excel pode ser usado para gerar diversas medidas de posição e de variabilidade de uma única variável e para gerar a covariância e o coeficiente de correlação como medidas da associação entre duas variáveis.

Como Usar Funções do Excel

O Excel oferece funções para calcular a média, a mediana, a moda, a variância da amostra e o desvio padrão. Ilustramos o uso das funções do Excel calculando a média, a mediana, a moda, a variância da amostra e o desvio padrão dos dados de salários iniciais da Tabela 3.1. Consulte a Figura 3.13 à medida que descrevermos as etapas envolvidas. Os dados estão inseridos na coluna B.

ARQUIVO DA INTERNET
Salary

A função MÉDIA do Excel pode ser utilizada para calcular a média ao digitarmos a seguinte fórmula na célula E1:

=MÉDIA(B2:B13)

Similarmente, as fórmulas =MED(B2:B13), =MODO(B2:B13), =VAR(B2:B13) e =DESVPAD(B2:B13) são inseridas nas células E2:E15, respectivamente, para calcular a mediana, a moda, a variância e o desvio padrão.

Figura 3.13 O uso de funções do Excel para calcular a média, a mediana, a moda, a variância e o desvio padrão

	A	B	C	D	E	F
1	Graduados	Salário Inicial		Média	=MÉDIA(B2:B13)	
2	1	2850		Mediana	=MED(B2:B13)	
3	2	2950		Moda	=MODO(B2:B13)	
4	3	3050		Variância	=VAR(B2:B13)	
5	4	2880		desvio padrão	=DESVPAD(B2:B13)	
6	5	2755				
7	6	2710				
8	7	2890				
9	8	3130				
10	9	2940				
11	10	3325				
12	11	2920				
13	12	2880				
14						

	A	B	C	D	E	F
1	Graduados	Salário Inicial		Média	2940	
2	1	2850		Mediana	2905	
3	2	2950		Moda	2880	
4	3	3050		Variância	27440.91	
5	4	2880		Desvio Padrão	165.65	
6	5	2755				
7	6	2710				
8	7	2890				
9	8	3130				
10	9	2940				
11	10	3325				
12	11	2920				
13	12	2880				
14						

A planilha que está em primeiro plano mostra que os valores computados usando-se funções do Excel são idênticos aos valores calculados anteriormente neste capítulo.

O Excel também provê funções que podem ser usadas para calcular a covariância e o coeficiente de correlação. Devemos ser cautelosos ao usar essas funções, uma vez que a função de covariância trata os dados como se estes fossem uma população, e a função de correlação trata os dados como uma amostra. Desse modo, o resultado obtido usando-se a função de covariância do Excel deve ser ajustado de forma que forneça a covariância da amostra. Mostramos aqui como essas funções podem ser usadas para calcular a covariância da amostra e o coeficiente de correlação da amostra dos dados da loja de equipamentos de som da Tabela 3.7. Consulte a Figura 3.14 à medida que apresentarmos as etapas envolvidas.

ARQUIVO DA INTERNET
Stereo

A função de covariância do Excel, COVAR, pode ser usada para calcular a covariância da população ao digitarmos a seguinte fórmula na célula F1:

=COVAR(B2:B11,C2:C11)

Similarmente, a fórmula =CORREL(B2:B11,C2:C11) é inserida na célula F2 para calcular o coeficiente de correlação da amostra. A planilha apresentada em segundo plano mostra os valores calculados usando-se as funções do Excel. Observe que o valor do coeficiente de correlação da amostra (0,93) é idêntico ao calculado usando-se a Equação (3.12). No entanto, o resultado 9,9 produzido pela função COVAR do Excel foi obtido tratando-se os dados como uma população. Desse modo, precisamos ajustar o resultado 9,9 do Excel para obtermos a covariância da amostra. O ajuste é bastante simples. Primeiramente, note que a fórmula para a covariância da população, a Equação (3.11), exige uma divisão pelo número total de observações no conjunto de dados. Mas a fórmula para a covariância da amostra, a Equação (3.10), exige uma divisão pelo número total de observações menos 1.

Figura 3.14 O uso de funções do Excel para calcular a covariância e a correlação

	A	B	C	D	E	F	G
1	Semana	Comerciais	Vendas		Covariância da População	=COVAR(B2:B11:C2:C11)	
2	1	2	50		Correlação da Amostra	=CORREL(B2:B11,C2:C11)	
3	2	5	57				
4	3	1	41				
5	4	3	54				
6	5	4	54				
7	6	1	38				
8	7	5	63				
9	8	3	48				
10	9	4	59				
11	10	2	46				
12							

	A	B	C	D	E	F	G
1	Semana	Comerciais	Vendas		Covariância da População	9,90	
2	1	2	50		Correlação da Amostra	0,93	
3	2	5	57				
4	3	1	41				
5	4	3	54				
6	5	4	54				
7	6	1	38				
8	7	5	63				
9	8	3	48				
10	9	4	59				
11	10	2	46				
12							

Então, para usarmos o resultado do Excel, 9,9, para calcular a covariância da amostra, simplesmente multiplicamos 9,9 por $n/(n-1)$. Uma vez que $n = 10$, obtemos;

$$s_{xy} = \left(\frac{10}{9}\right)9{,}9 = 11$$

Assim, a covariância da amostra para os dados da loja de equipamentos de som é 11.

Como Usar a Ferramenta Estatística Descritiva do Excel

Conforme já demonstramos, o Excel oferece funções estatísticas para calcular estatísticas descritivas de um conjunto de dados. Essas funções podem ser usadas para calcular uma estatística a cada vez (por exemplo, a média, a variância etc.). O Excel também oferece uma série de Ferramentas de Análise de Dados. Uma dessas ferramentas, cujo nome é Estatística Descritiva, permite ao usuário calcular uma série de estatísticas descritivas simultaneamente. Mostramos aqui como ela pode ser usada para calcular as estatísticas descritivas dos dados de salários iniciais apresentados na Tabela 3.1. Consulte a Figura 3.15 à medida que descrevermos as etapas envolvidas.

ARQUIVO DA INTERNET

Salary

Etapa 1. Selecione o menu **Ferramentas**

Etapa 2. Escolha **Análise de Dados**

Etapa 3. Quando a caixa de diálogo Análise de Dados aparecer:
- Escolha **Estatística Descritiva**
- Dê um clique em **OK**

Etapa 4. Quando a caixa de diálogo Estatística Descritiva aparecer:
- Digite B1:B13 na caixa **Intervalo de Entrada**
- Selecione **Agrupado por Colunas**
- Selecione **Rótulos na Primeira Linha**
- Selecione **Intervalo de Saída**
- Digite D1 na caixa **Intervalo de Saída** (para identificar o canto superior esquerdo da parte da planilha em que a estatística descritiva aparecerá)
- Selecione **Resumo Estatístico**
- Dê um clique em **OK**

Figura 3.15 Dados de saída (*output*) da ferramenta estatística descritiva do Excel

	A	B	C	D	E	F
1	**Graduados**	**Salário Inicial**		***Salário Inicial***		
2	1	2850				
3	2	2950		**Média**	2940	
4	3	3050		Erro Padrão	47.82	
5	4	2880		**Mediana**	2905	
6	5	2755		**Moda**	2880	
7	6	2710		**Desvio Padrão**	165,65	
8	7	2890		**Variância da Amostra**	27440,91	
9	8	3130		Achatamento	1.7189	
10	9	2940		**Assimetria**	1.0911	
11	10	3325		**Amplitude**	615	
12	11	2920		**Mínimo**	2710	
13	12	2880		**Máximo**	3325	
14				**Soma**	35280	
15				**Contagem**	12	
16						

As células D1:E15 da Figura 3.15 apresentam a estatística descritiva produzida pelo Excel. As entradas em negrito são as estatísticas descritivas que abordamos neste capítulo. As estatísticas descritivas que não estão em negrito ou serão abordadas posteriormente ou serão discutidas mais detalhadamente ao longo do livro.

Introdução à Probabilidade

ESTATÍSTICA NA PRÁTICA

MORTON INTERNATIONAL*
Chicago, Illinois

A Morton International é uma empresa que comercializa sal, produtos domésticos, motores de foguetes e química fina. A Carstab Corporation, uma subsidiária da Morton International, produz química fina e disponibiliza uma série de produtos químicos concebidos para cumprir as especificações exclusivas de seus clientes. Para um cliente em particular a Carstab produziu um custoso catalisador que é usado no processamento de produtos químicos. Alguns lotes, mas não todos, produzidos pela Carstab satisfazem as especificações do cliente para o produto.

O cliente da Carstab concordou em testar cada lote depois de recebê-lo e determinar se o catalisador desempenharia a função desejada. Os lotes que não fossem aprovados no teste realizado pelo cliente seriam devolvidos à Carstab. No decorrer do tempo, a Carstab descobriu que o cliente aceitava 60% dos lotes e devolvia 40%. Em termos de probabilidade, cada remessa da Carstab ao cliente tinha uma probabilidade de 0,60 de ser aceita e uma probabilidade de 0,40 de ser devolvida.

Nem a Carstab nem seu cliente estavam satisfeitos com esses resultados. Em um esforço para melhorar o serviço, a Carstab explorou a possibilidade de reproduzir o teste do cliente antes do embarque. Entretanto, o alto custo dos equipamentos especiais de teste tornou inviável essa alternativa. Os químicos da Carstab pro-

* Os autores agradecem a Michael Haskell, da Morton International, por fornecer esta "Estatística na Prática".

puseram então um novo teste de custo relativamente baixo idealizado para indicar se um lote seria aprovado no teste do cliente. A questão de probabilidade envolvida era: qual a probabilidade de um lote ser aprovado no teste do cliente se tivesse sido aprovado no novo teste da Carstab?

Uma amostra de lotes foi produzida e submetida ao novo teste da Carstab. Somente os lotes aprovados no novo teste eram enviados ao cliente. A análise probabilística dos dados indicou que, se um lote fosse aprovado no teste da Carstab, teria uma probabilidade de 0,909 de ser aprovado no teste do cliente e ser aceito. Alternativamente, se um lote fosse aprovado no teste da Carstab, teria somente uma probabilidade de 0,091 de ser devolvido. A análise probabilística forneceu uma comprovação fundamental para a adoção e implementação dos novos procedimentos de teste na Carstab. O novo teste resultou em uma melhoria imediata do atendimento ao cliente e em uma redução substancial dos custos de embarque e manuseio dos lotes devolvidos.

A probabilidade de um lote ser aceito pelo cliente depois de ser aprovado no novo teste da Carstab denomina-se probabilidade condicional. Neste capítulo, você aprenderá a calcular esta e outras probabilidades que são úteis no processo de tomada de decisões.

Alguns dos primeiros trabalhos sobre probabilidade originaram-se de uma série de cartas trocadas entre Pierre de Fermat e Blaise Pascal nos idos de 1650.

Os gerentes freqüentemente fundamentam suas decisões em uma análise de incertezas, como as que apresentamos a seguir:

1. Quais são as chances de queda das vendas se aumentarmos os preços?
2. Qual é a probabilidade de um novo método de montagem aumentar a produtividade?
3. Qual é a probabilidade de o projeto ser concluído no prazo?
4. Qual é a chance de um novo investimento ser lucrativo?

Probabilidade é uma medida numérica da possibilidade de um evento ocorrer. Desse modo, podemos usar probabilidades como medidas do grau de incerteza associado aos quatro eventos anteriormente relacionados. Se houver probabilidades disponíveis, podemos determinar a possibilidade de cada um dos eventos ocorrer.

Valores probabilísticos sempre são atribuídos em uma escala de 0 a 1. Uma probabilidade próxima de 0 indica que é improvável que um evento ocorra; uma probabilidade próxima de 1 revela que a ocorrência de um evento é quase certa.

Outras probabilidades entre 0 e 1 representam o grau de possibilidade de um evento vir a ocorrer. Por exemplo, se considerarmos o evento "chover amanhã", entendemos que, quando o boletim meteorológico indica "uma probabilidade de chuva próxima de zero", isso quer dizer que não há quase chance alguma de chover. Entretanto, se houver a indicação de 0,90 de probabilidade de chuva, saberemos que é provável que ocorra chuva. Uma probabilidade de 0,50 mostra que tanto é possível chover como não. A Figura 4.1 retrata a imagem da probabilidade como uma medida numérica da possibilidade de um evento ocorrer.

4.1 EXPERIMENTOS, REGRAS DE CONTAGEM E ATRIBUINDO PROBABILIDADES

Ao discutirmos a probabilidade, definimos um experimento como um processo que gera resultados bem definidos. Em uma única repetição de um experimento, ocorrerá um, e somente um, dos resultados experimentais possíveis. Diversos exemplos de experimentos e seus respectivos resultados são apresentados a seguir:

Experimento	Resultados Experimentais
Jogar uma moeda	Cara, coroa
Selecionar uma peça para inspeção	Defeituosa, não-defeituosa
Fazer um contato de vendas	Comprar, não comprar
Lançar um dado	1, 2, 3, 4, 5, 6
Jogar uma partida de futebol	Ganhar, perder, empatar

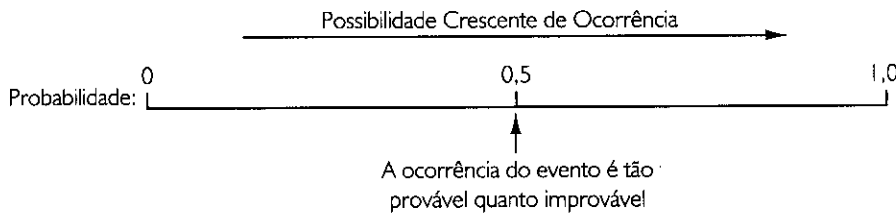
Ao especificar todos os resultados possíveis, identificamos o **espaço amostral** de um experimento.

ESPAÇO AMOSTRAL

O espaço amostral de um experimento é o conjunto de todos os resultados experimentais.

Um resultado experimental também é chamado **ponto amostral** para identificá-lo como um elemento do espaço amostral.

Resultados experimentais são também denominados pontos amostrais.

Figura 4.1 A probabilidade como uma medida numérica da possibilidade de ocorrência de um evento

Considere o primeiro experimento da tabela anterior – jogar uma moeda. A face da moeda voltada para cima – cara ou coroa – determina os resultados experimentais (pontos amostrais). Dado que S denota o espaço amostral, podemos usar a seguinte notação para descrever o espaço amostral:

$$S = \{\text{Cara, Coroa}\}$$

O espaço amostral do segundo experimento da tabela – selecionar uma peça para inspeção – pode ser descrito da seguinte maneira:

$$S = \{\text{Defeituoso, Não defeituoso}\}$$

Ambos os experimentos que acabamos de descrever têm dois resultados experimentais (pontos amostrais). Entretanto, suponha que consideremos o quarto experimento relacionado na tabela: lançar um dado. Os resultados experimentais possíveis, definidos como o número de pontos que aparecem na face superior do dado, são os seis pontos do espaço amostral desse experimento:

$$S = \{1, 2, 3, 4, 5, 6\}$$

Regras de Contagem, Combinações e Permutações

Ser capaz de identificar e contar os resultados amostrais é uma etapa necessária na atribuição de probabilidades. Vamos discutir agora três regras de contagem úteis.

Experimentos em múltiplas etapas. A primeira regra de contagem aplica-se a experimentos que são feitos em múltiplas etapas. Considere o experimento de jogar duas moedas. Digamos que os resultados sejam definidos em termos do padrão de caras e coroas que aparecem nas faces voltadas para cima das duas moedas. Quantos resultados experimentais são possíveis para esse experimento? O experimento de jogar duas moedas pode ser imaginado como um experimento de duas etapas no qual a etapa 1 consiste em lançar a primeira moeda, e a etapa 2, em lançar a segunda moeda. Se usarmos H para denotar cara e T, para coroa (H, H), isso indicará o espaço experimental com cara na primeira moeda e coroa na segunda moeda. Prosseguindo com essa notação, podemos descrever o espaço amostral (S) desse experimento de lançar a moeda da seguinte maneira:

$$S = \{(H, H), (H, T), (T, H), (T, T)\}$$

Desse modo, notamos que quatro resultados experimentais são possíveis. Nesse caso, podemos enumerar facilmente todos os resultados experimentais.

A regra de contagem de experimentos em múltiplas etapas torna possível determinar o número de resultados experimentais sem enumerá-los.

REGRA DE CONTAGEM DE EXPERIMENTOS EM MÚLTIPLAS ETAPAS

Se um experimento pode ser descrito como uma seqüência de k etapas com n_1 resultados possíveis na primeira etapa, n_2 resultados possíveis na segunda etapa e assim por diante, o número total de resultados experimentais será dado por $(n_1)\,(n_2)\ldots(n_k)$.

Considerando o experimento de lançar duas moedas como uma seqüência de lançar primeiro uma moeda ($n_1 = 2$) e depois lançar a outra moeda ($n_2 = 2$), podemos ver a partir da regra de contagem que há (2)(2) = 4 resultados experimentais distintos. Conforme mostramos anteriormente, eles são $S = \{(H, H), (H, T), (T, H), (T, T)\}$. O número de resultados em um experimento que envolve lançar seis moedas é (2)(2)(2)(2)(2)(2) = 64.

Sem o diagrama em árvore poder-se-ia pensar que somente três resultados experimentais são possíveis para dois lançamentos de uma moeda: nenhuma cara, 1 cara e 2 caras.

Um **diagrama em árvore** é uma representação gráfica que ajuda a visualizar um experimento em múltiplas etapas. A Figura 4.2 mostra um diagrama em árvore correspondente ao experimento de lançar duas moedas. A seqüência de etapas desloca-se da esquerda para a direita ao longo do diagrama. A etapa 1 corresponde ao lançamento da primeira moeda, e a etapa 2 refere-se ao lançamento da segunda moeda. Para cada etapa, os dois resultados possíveis são cara ou coroa. Note que, para cada resultado possível na etapa 1, há duas ramificações que correspondem aos dois resultados possíveis na etapa 2. Cada um dos pontos no lado direito da árvore referentes aos dois resultados possíveis corresponde a um resultado experimental. Cada percurso ao longo da árvore, do nó localizado na extremidade esquerda a um dos nós no lado direito da árvore, corresponde a uma seqüência individual de resultados.

Vejamos agora como a regra de contagem de experimentos em múltiplas etapas pode ser usada na análise de um projeto de ampliação da capacidade na Kentucky Power & Light Company (KP&L). A KP&L está iniciando um projeto idealizado para aumentar a capacidade de geração de energia em uma de suas usinas ao norte de Kentucky. O projeto divide-se em duas etapas, ou passos, seqüenciais: etapa 1 (projeto) e etapa 2 (construção). Não obstante cada etapa estar programada e ser controlada o mais cuidadosamente possível, a administração não é capaz de prever o tempo exato necessário para o término de cada fase do projeto. Uma análise de projetos de construção similares revelou que os prazos de término possíveis para a fase de elaboração do projeto seriam 2, 3 ou 4 meses, e que os prazos de término para a fase de construção seriam 6, 7 ou 8 meses. Além disso, em virtude da necessidade crítica de energia elétrica adicional, a administração estabeleceu uma meta de dez meses para a conclusão total do projeto.

Desde que esse projeto tem três prazos de término possíveis para a fase de elaboração do projeto (etapa 1) e três prazos de término possíveis para a fase de construção (etapa 2), a regra de contagem para experimentos em múltiplas etapas pode ser aplicada nesse caso para determinar um total de (3)(3) = 9 resultados experimentais. Para descrever os resultados experimentais, usaremos uma notação de dois números: por exemplo, (2, 6) indica que a fase de projeto será concluída em dois meses e a fase de construção, em 6 meses. Esse resultado experimental representa um total de 2 + 6 = 8 meses para a conclusão total do projeto. A Tabela 4.1 sintetiza os nove resultados experimentais para o problema da KP&L. O diagrama em árvore da Figura 4.3 mostra como ocorrem os nove resultados (pontos amostrais).

A regra de contagem e o diagrama em árvore ajudam o gerente de projetos a identificar os resultados experimentais e determinar os prazos possíveis para o término do projeto. A partir da informação da Figura 4.3, notamos que o projeto será concluído em um prazo de oito a 12 meses, com seis dos nove resultados experimentais apresentando o prazo de conclusão desejado de dez meses ou menos.

Figura 4.2 Diagrama em árvore do experimento de lançar duas moedas

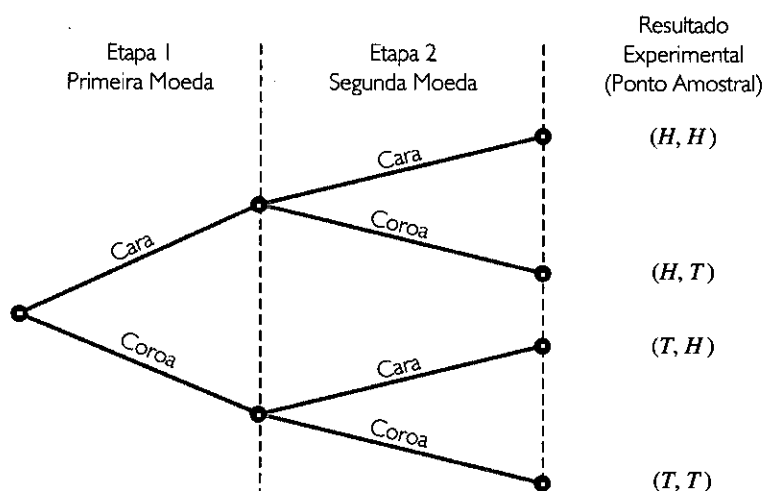
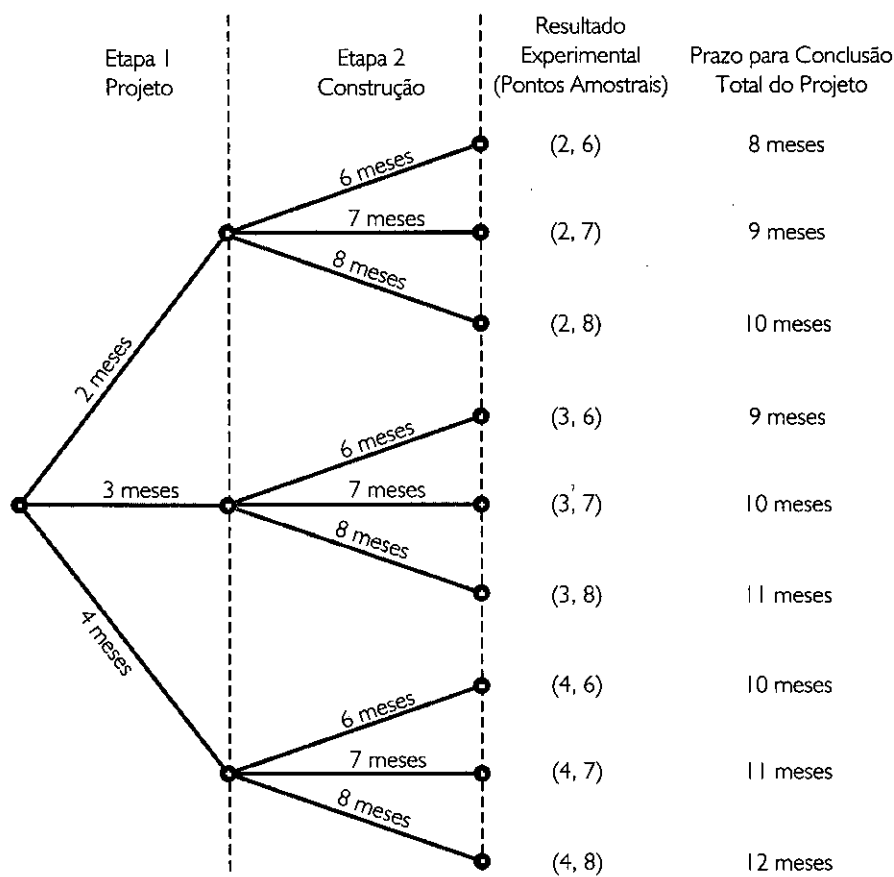



Tabela 4.1 Resultados experimentais (pontos amostrais) correspondentes ao projeto da KP&L

Prazo de Término (meses)			
Etapa 1 Elaboração do Projeto	Etapa 2 Construção	Notação para o Resultado Experimental	Prazo para Conclusão Total do Projeto (meses)
2	6	(2, 6)	8
2	7	(2, 7)	9
2	8	(2, 8)	10
3	6	(3, 6)	9
3	7	(3, 7)	10
3	8	(3, 8)	11
4	6	(4, 6)	10
4	7	(4, 7)	11
4	8	(4, 8)	12

Figura 4.3 Diagrama em árvore do projeto da KP&L

Embora a identificação dos resultados experimentais possa ser útil, precisamos considerar como podemos atribuir valores probabilísticos aos resultados experimentais antes de fazer uma avaliação da probabilidade de que o projeto venha a ser concluído dentro do prazo desejado de dez meses.

Combinações. Uma segunda regra útil de contagem nos permite contar o número de resultados experimentais quando o experimento envolve escolher n objetos de um conjunto (geralmente maior) de N objetos. Ela se denomina regra de contagem de combinações.

REGRA DE CONTAGEM DE COMBINAÇÕES

O número de combinações de N objetos, tomados n a cada vez, é:

$$C_n^N = \binom{N}{n} = \frac{N!}{n!(N-n)!} \tag{4.1}$$

em que

$$N! = N(N-1)(N-2)\cdots(2)(1)$$
$$n! = n(n-1)(n-2)\cdots(2)(1)$$

e, por definição,

$$0! = 1$$

A notação ! significa *fatorial*; por exemplo, 5 fatorial é $5! = (5)(4)(3)(2)(1) = 120$.

Como ilustração da regra de contagem de combinações, considere um procedimento de controle da qualidade em que um inspetor seleciona aleatoriamente duas de cinco peças para testar se há defeitos. Em um grupo de cinco peças, quantas combinações de duas peças podem ser selecionadas? A regra de contagem da Equação 4.1 mostra que, com $N = 5$ e $n = 2$, teremos:

Na amostragem de uma população finita de tamanho N, usamos a regra de contagem de combinações para encontrar o número de diferentes amostras de tamanho n que podem ser selecionadas.

$$C_2^5 = \binom{5}{2} = \frac{5!}{2!(5-2)!} = \frac{(5)(4)(3)(2)(1)}{(2)(1)(3)(2)(1)} = \frac{120}{12} = 10$$

Desse modo, dez resultados são possíveis para o experimento de escolher aleatoriamente duas peças de um grupo de cinco. Se rotularmos as cinco peças como A, B, C, D e E, as dez combinações ou resultados experimentais podem ser identificados como AB, AC, AD, AE, BC, BD, BE, CD, CE e DE.

Como outro exemplo, considere que o sistema lotérico de Ohio utilize a escolha aleatória de seis números inteiros de um grupo de 47 para determinar o ganhador da loteria semanal. A regra de contagem de combinações, Equação 4.1, pode ser usada para determinar o número de maneiras pelas quais os seis diferentes números inteiros podem ser escolhidos de um grupo de 47.

$$\binom{47}{6} = \frac{47!}{6!(47-6)!} = \frac{47!}{6!41!} = \frac{(47)(46)(45)(44)(43)(42)}{(6)(5)(4)(3)(2)(1)} = 10.737.573$$

A regra de contagem de combinações nos diz que mais de 10 milhões de resultados experimentais são possíveis no sorteio da loteria. Uma pessoa que compra um bilhete dessa loteria tem uma chance em 10.737.573 de ganhar.

A regra de contagem de combinações mostra que a chance de ganhar na loteria é muito improvável.

Permutações. Uma terceira regra de contagem que às vezes é útil é a regra de contagem de permutações. Ela permite a uma pessoa calcular o número de resultados experimentais quando n objetos são escolhidos de um conjunto de N objetos em que a ordem de escolha é importante. Os mesmos n objetos escolhidos em uma ordem diferente são considerados um resultado experimental diferente.

REGRA DE CONTAGEM DE PERMUTAÇÕES

O número de permutações de N objetos, tomados n a cada vez, é dado por:

$$P_n^N = n!\binom{N}{n} = \frac{N!}{(N-n)!} \tag{4.2}$$

A regra de contagem de permutações está estreitamente relacionada com a das combinações; entretanto, um experimento resulta em mais permutações do que combinações para o mesmo número de objetos porque cada escolha de n objetos pode ser organizada em $n!$ maneiras diferentes.

Como exemplo, considere novamente o processo de controle da qualidade no qual o inspetor escolhe duas de cinco peças para inspecioná-las à procura de defeitos. Quantas permutações podem ser escolhidas? A regra de contagem da Equação 4.2 mostra que com $N = 5$ e $n = 2$, teremos:

$$P_2^5 = \frac{5!}{(5-2)!} = \frac{5!}{3!} = \frac{(5)(4)(3)(2)(1)}{(3)(2)(1)} = \frac{120}{6} = 20$$

Desse modo, são possíveis 20 resultados para o experimento de escolher aleatoriamente duas peças de um grupo de cinco quando a ordem de escolha deve ser levada em consideração. Se rotularmos as peças como A, B, C, D e E, as 20 permutações serão: AB, BA, AC, CA, AD, DA, AE, EA, BC, CB, BD, DB, BE, EB, CD, DC, CE, EC, DE e ED.

Atribuindo Probabilidades

Vejamos agora como se pode atribuir probabilidades a resultados experimentais. As três abordagens usadas com maior freqüência são o método clássico, o de freqüência relativa e o subjetivo. Independentemente do método utilizado, dois **requisitos básicos para atribuição de probabilidades** devem ser satisfeitos:

REQUISITOS BÁSICOS PARA ATRIBUIÇÃO DE PROBABILIDADES

1. A probabilidade atribuída a cada um dos resultados experimentais deve situar-se entre 0 e 1, inclusive. Se admitirmos que E_i denota o i-ésimo resultado experimental e que $P(E_i)$ é sua probabilidade, então esse requisito pode ser escrito na seguinte forma:

$$0 \leq P(E_i) \leq 1 \text{ para todo } i \qquad \textbf{(4.3)}$$

2. A soma das probabilidades de todos os resultados experimentais deve ser igual a 1,0. Para n resultados experimentais, esse requisito pode ser escrito na seguinte forma:

$$P(E_1) + P(E_2) + \ldots + P(E_n) = 1 \qquad \textbf{(4.4)}$$

O **método clássico** de atribuição de probabilidades é apropriado quando todos os resultados experimentais são igualmente prováveis. Se n resultados experimentais são possíveis, a probabilidade de $1/n$ é atribuída a cada resultado experimental. Quando se usa essa abordagem, os dois requisitos para atribuição de probabilidade são automaticamente satisfeitos.

Como exemplo, considere o experimento de jogar uma moeda; os dois resultados experimentais – cara e coroa – são igualmente prováveis. Uma vez que um dos dois resultados igualmente prováveis é cara, a probabilidade de se observar cara é 1/2, ou 0,50. De forma similar, a probabilidade de se observar coroa também é 1/2, ou 0,50.

Como outro exemplo, considere o experimento de lançar um dado. Seria razoável concluirmos que os seis resultados possíveis são igualmente prováveis e, portanto, a cada resultado é atribuída uma probabilidade de 1/6. Se $P(1)$ denota a probabilidade de 1 aparecer na face do dado voltada para cima, então $P(1) = 1/6$. Similarmente, $P(2) = 1/6$, $P(3) = 1/6$, $P(4) = 1/6$, $P(5) = 1/6$ e $P(6) = 1/6$. Observe que essas probabilidades satisfazem os dois requisitos básicos das Equações 4.3 e 4.4 porque cada uma das probabilidades é maior ou igual a zero e sua soma é 1,0.

O **método de freqüência relativa** para a atribuição de probabilidades é apropriado quando se tem dados disponíveis para estimar a proporção do tempo em que o resultado experimental ocorrerá se o experimento for repetido inúmeras vezes. Como exemplo, considere um estudo sobre o tempo de espera no setor de raios X de um hospital municipal. Um atendente registrou o número de pacientes à espera de atendimento às 9h em 20 dias consecutivos e obteve os seguintes resultados:

Número de Pessoas a Espera	Número de Dias em que o Resultado Ocorreu
0	2
1	5
2	6
3	4
4	3
Total	20

Esses dados mostram que em dois dos 20 dias, nenhum (0) paciente estava à espera de atendimento; em cinco desses dias, um paciente estava à espera de atendimento e assim por diante. Usando o método de freqüência relativa, atribuiríamos uma probabilidade de 2/20 = 0,10 ao resultado experimental de nenhum paciente estar à espera de atendimento, 5/20 = 0,25 ao resultado experimental de um paciente estar à espera, 6/20 = 0,30 para dois pacientes, 4/20 = 0,20 para três pacientes e 3/20 = 0,15 para quatro

pacientes à espera. A exemplo do que ocorre com o método clássico, usar o método de freqüência relativa satisfaz automaticamente os dois requisitos básicos das Equações 4.3 e 4.4.

O **método subjetivo** de atribuição de probabilidades é o mais apropriado quando não se pode presumir realisticamente que os resultados experimentais são igualmente prováveis e quando poucos dados relevantes estão disponíveis. Quando o método subjetivo é usado para atribuir probabilidades aos resultados experimentais, podemos usar qualquer informação disponível, como nossa experiência ou intuição. Depois de considerarmos todas as informações disponíveis, especificamos um valor probabilístico que expresse nosso *grau de confiança* (em uma escala de 0 a 1) de que o resultado experimental ocorrerá. Quando se usa o método subjetivo, pode-se esperar que diferentes pessoas atribuam diferentes probabilidades ao mesmo resultado experimental.

O método subjetivo exige que se tenha um cuidado extra para assegurar que os dois requisitos básicos das Equações 4.3 e 4.4 sejam satisfeitos. Independentemente do grau de confiança de uma pessoa, o valor probabilístico atribuído a cada resultado experimental deve situar-se entre 0 e 1, inclusive, e a soma de todas as probabilidades para os resultados experimentais deve ser igual a 1,0.

Considere o caso em que Tom e Judy Elsbernd fizeram uma oferta para comprar uma casa. São dois os resultados possíveis:

$$E_1 = \text{sua oferta é aceita}$$
$$E_2 = \text{sua oferta é rejeitada}$$

Judy acredita que a probabilidade de sua oferta ser aceita é 0,8; assim, Judy estabeleceria que $P(E_1) = 0{,}8$ e $P(E_2) = 0{,}2$. Tom, entretanto, acredita que a probabilidade de sua oferta ser aceita é 0,6; portanto, Tom estabeleceria que $P(E_1) = 0{,}6$ e $P(E_2) = 0{,}4$. Note que a estimativa de probabilidade de Tom para E_1 reflete um pessimismo maior quanto à possibilidade de que sua oferta seja aceita.

Tanto Tom como Judy atribuíram probabilidades que satisfazem os dois requisitos básicos. O fato de suas estimativas de probabilidade serem diferentes enfatiza a natureza pessoal do método subjetivo.

O teorema de Bayes (veja a Seção 4.5) constitui uma maneira de combinarmos probabilidades anteriores, as quais são determinadas subjetivamente, com probabilidades obtidas por outros meios, a fim de obtermos probabilidades revistas ou posteriores.

Mesmo em situações de negócios em que a abordagem clássica ou a de freqüência relativa podem ser aplicadas, os gerentes podem querer produzir estimativas de probabilidade subjetivas. Nesses casos, as melhores estimativas de probabilidade freqüentemente são obtidas combinando-se as estimativas obtidas da abordagem clássica ou de freqüência relativa com as estimativas de probabilidade subjetivas.

Probabilidades do Projeto da KP&L

Para realizarmos uma análise adicional do projeto da KP&L precisamos desenvolver probabilidades para cada um dos nove resultados experimentais relacionados na Tabela 4.1. Com base na experiência e na capacidade de julgamento, a administração concluiu que os resultados experimentais não eram igualmente prováveis. Portanto, o método clássico de atribuição de probabilidades não poderia ser usado. A administração decidiu então realizar um estudo dos prazos de conclusão de projetos similares levados a efeito pela KP&L ao longo dos três últimos anos. Os resultados de um estudo de 40 projetos similares estão resumidos na Tabela 4.2.

Depois de rever os resultados do estudo, a administração decidiu empregar o método de freqüência relativa de atribuição de probabilidades. A administração poderia ter produzido estimativas de probabilidade subjetivas, mas achou que o projeto atual era muito similar aos 40 projetos anteriores. Desse modo, o método de freqüência relativa foi considerado o melhor.

Ao usar os dados da Tabela 4.2 para calcular as probabilidades, observamos que o resultado (2, 6) – ou seja, a etapa 1 concluída em dois meses e a etapa 2 concluída em seis meses – ocorria seis vezes nos 40 projetos. Podemos usar o método de freqüência relativa para atribuir uma probabilidade de $6/40 = 0{,}15$ a esse resultado. De forma similar, o resultado (2, 7) também ocorreu em seis dos 40 projetos. Produzindo uma probabilidade de $6/40 = 0{,}15$. Prosseguindo dessa maneira, obtemos as atribuições de probabilidade para os pontos amostrais do projeto da KP&L mostrados na Tabela 4.3. Note que $P(2, 6)$ representa a probabilidade do ponto amostral (2, 6), $P(2, 7)$ representa a probabilidade do ponto amostral (2, 7) e assim por diante.

Tabela 4.2 Resultados do estudo relativo ao prazo de término de 40 projetos da KP&L

Prazo de Término (meses) Etapa 1 Elaboração do Projeto	Etapa 2 Construção	Ponto Amostral	Número de Projetos Anteriores que Tiveram Estes Prazos de Término
2	6	(2, 6)	6
2	7	(2, 7)	6
2	8	(2, 8)	2
3	6	(3, 6)	4
3	7	(3, 7)	8
3	8	(3, 8)	2
4	6	(4, 6)	2
4	7	(4, 7)	4
4	8	(4, 8)	6
		Total	40

Tabela 4.3 Atribuindo probabilidade para o projeto da KP&L com base no método da freqüência relativa

Ponto Amostral	Prazo de Término do Projeto	Probabilidade do Ponto Amostral
(2, 6)	8 meses	P(2, 6) 5 6/40 5 0,15
(2, 7)	9 meses	P(2, 7) 5 6/40 5 0,15
(2, 8)	10 meses	P(2, 8) 5 2/40 5 0,05
(3, 6)	9 meses	P(3, 6) 5 4/40 5 0,10
(3, 7)	10 meses	P(3, 7) 5 8/40 5 0,20
(3, 8)	11 meses	P(3, 8) 5 2/40 5 0,05
(4, 6)	10 meses	P(4, 6) 5 2/40 5 0,05
(4, 7)	11 meses	P(4, 7) 5 4/40 5 0,10
(4, 8)	12 meses	P(4, 8) 5 6/40 5 0,15
		Total 1,00

NOTAS E COMENTÁRIOS

1. Em estatística, a noção de experimento difere consideravelmente da noção de experimento nas ciências físicas. Nas ciências físicas, os pesquisadores geralmente realizam o experimento em um laboratório ou em um ambiente controlado a fim de conhecerem a causa e o efeito. Em experimentos estatísticos, a probabilidade determina os resultados. Não obstante o experimento ser repetido da mesma maneira, um resultado completamente diferente pode ocorrer. Em razão dessa influência da probabilidade sobre o resultado, os experimentos de estatística às vezes são chamados *experimentos aleatórios*.
2. Quando extraímos uma amostra aleatória, sem substituição, de uma população de tamanho N, usamos a regra de contagem de combinações para encontrar o número de diferentes amostras de tamanho n que podem ser selecionadas.

Exercícios

Métodos

AUTOTESTE

1. Um experimento tem três etapas com três resultados possíveis para a primeira etapa, dois resultados possíveis para a segunda etapa e quatro resultados possíveis para a terceira etapa. Quantos resultados experimentais existem para o experimento como um todo?
2. De quantas maneiras três itens podem ser selecionados de um grupo de seis itens? Use as letras A, B, C, D e E para identificar os itens e relacione cada uma das diferentes combinações dos três itens.
3. Quantas permutações de três itens podem ser selecionadas de um grupo de seis? Use as letras A, B, C, D e E para identificar os itens e relacione cada uma das permutações dos itens B, D e F.

4. Considere o experimento de lançar uma moeda três vezes.
   a. Desenvolva um diagrama em árvore para o experimento.
   b. Relacione os resultados experimentais.
   c. Qual é a probabilidade relativa a cada resultado experimental?
5. Suponha que um experimento tenha cinco resultados igualmente prováveis: E_1, E_2, E_3, E_4, E_5. Atribua probabilidades a cada resultado e demonstre que os requisitos indicados nas Equações 4.3 e 4.4 foram satisfeitos. Qual método você usou?

AUTOTESTE

6. Um experimento com três resultados foi repetido 50 vezes e soube-se que E_1 ocorria 20 vezes; E_2, 13 vezes; e E_3, 17 vezes. Atribua probabilidades aos resultados. Qual método você usou?
7. Um tomador de decisões atribuiu subjetivamente as seguintes probabilidades aos quatro resultados de um experimento: $P(E_1) = 0{,}10$, $P(E_2) = 0{,}15$, $P(E_3) = 0{,}40$ e $P(E_4) = 0{,}20$. Essas atribuições de probabilidade são válidas? Explique.

Aplicações

8. Na cidade de Milford, os requerimentos para alteração do zoneamento passam por duas etapas: uma revisão pela comissão de planejamento e uma decisão da Câmara Municipal. Na etapa 1, a comissão de planejamento revisa o requerimento de alteração do zoneamento e apresenta uma recomendação positiva ou negativa correspondente. Na etapa 2, a Câmara Municipal revisa a recomendação da comissão de planejamento e então realiza uma votação para aprovar ou desaprovar a alteração do zoneamento. Considere o processo de requerimento um experimento.

   a. Quantos pontos amostrais há para esse experimento? Relacione-os.
   b. Construa um diagrama em árvore para o experimento.

AUTOTESTE

9. A amostragem aleatória simples usa uma amostra de tamanho n de uma população de tamanho N para obter dados que podem ser usados para se fazer inferências a respeito das características de uma população. Suponha que de uma população de 50 contas bancárias queiramos extrair uma amostra aleatória de quatro contas a fim de conhecermos a população. Quantas amostras aleatórias diferentes de quatro contas são possíveis?

AUTOTESTE

10. O capital para investimento (*venture capital*) pode oferecer um grande impulso aos fundos disponíveis para as empresas. De acordo com a Venture Economics (*Investor's Business*, 28 de abril de 2000), dos 2.374 desembolsos de capital para investimento, 1.434 foram para empresas da Califórnia, 390 para empresas de Massachussetts, 217 para empresas de Nova York e 112 para empresas do Colorado. Vinte e dois por cento das empresas que recebem fundos se encontravam nas primeiras etapas de desenvolvimento e 55% das empresas, na fase de expansão. Suponha que você queira escolher aleatoriamente uma dessas empresas para saber como os fundos de capital para investimento são usados.

   a. Qual é a probabilidade de a empresa escolhida ser da Califórnia?
   b. Qual é a probabilidade de a empresa escolhida não ser de nenhum dos quatro estados mencionados?
   c. Qual é a probabilidade de a empresa não estar nas primeiras etapas de desenvolvimento?
   d. Supondo que as empresas que se encontravam nas primeiras etapas de desenvolvimento estavam uniformemente distribuídas pelo território nacional, quantas empresas de Massachussetts que recebem fundos de capital para investimento estavam nas primeiras etapas de desenvolvimento?
   e. A quantia total de fundos investidos foi de US$ 32,4 bilhões. Estime o valor que foi destinado ao Colorado.
11. A National Highway Traffic Safety Administration (NHTSA) realizou uma pesquisa para saber como os motoristas norte-americanos usam os cintos de segurança (Associated Press, 25 de agosto de 2003). Dados de amostra coerentes com a pesquisa realizada pela NHTSA são apresentados a seguir:

	O Motorista Usa o Cinto de Segurança?	
Região	**Sim**	**Não**
Nordeste	148	52
Meio-Oeste	162	54
Sul	296	74
Oeste	252	48
Total	858	228

a. Em relação aos Estados Unidos, qual é a probabilidade de um motorista usar o cinto de segurança?
b. A probabilidade de uso do cinto de segurança por um motorista norte-americano foi de 0,75. O diretor da NHTSA, Dr. Jeffrey Runge, esperava uma probabilidade de 0,78 para 2003. Ele teria ficado satisfeito com os resultados da pesquisa de 2003?
c. Qual é a probabilidade de uso do cinto de segurança de acordo com a região do país? Qual região apresenta o maior uso do cinto de segurança?
d. Qual proporção dos motoristas integrantes da amostra vieram de cada uma das regiões do país? Qual região teve o maior número de motoristas selecionados? Qual região teve o segundo maior número de motoristas selecionados?
e. Supondo que o número total de motoristas de cada região seja o mesmo, você vê alguma razão pela qual a estimativa probabilística do item (a) poderia ser demasiadamente elevada? Explique.

12. A loteria Powerball é jogada duas vezes por semana em 23 estados, nas Ilhas Virgens e no Distrito de Colúmbia. Para jogar na Powerball o participante deve comprar um bilhete que custa US$ 1 e então escolher cinco números dos dígitos 1 a 53 e um número Powerball dos dígitos 1 a 42. Para determinar os números dados em cada jogo, os diretores da loteria extraem cinco bolas brancas de um globo com 53 bolas brancas, e uma bola vermelha de um globo com 42 bolas vermelhas. Para ganhar o prêmio, os números do bilhete do participante devem coincidir com os números contidos nas cinco bolas brancas, em qualquer ordem, bem como o número Powerball. Em agosto de 2001, quatro ganhadores repartiram um prêmio de US$ 295 milhões ao acertarem os números 8 – 17 – 22 – 42 – 47, mais o número Powerball 21. Além do prêmio principal, há uma série de outros prêmios que são concedidos a cada vez que há sorteios. Por exemplo, um prêmio de US$ 100 mil é pago se os cinco números do participante coincidirem com os cinco números contidos nas cinco bolas brancas (www.powerball.com, 25 de março de 2003).
    a. Calcule o número de maneiras pelas quais os cinco primeiros números podem ser selecionados.
    b. Qual é a probabilidade de se ganhar um prêmio de US$ 100 mil ao coincidir os números contidos nas cinco bolas brancas?
    c. Qual é a probabilidade de se ganhar o prêmio Powerball?

13. Uma empresa que produz creme dental estuda cinco diferentes desenhos (*designs*) de embalagem. Supondo que um desenho tenha exatamente a mesma probabilidade de ser escolhido pelo cliente que outro qualquer, qual probabilidade de escolha você atribuiria a cada um dos desenhos de embalagem? Em um experimento real, 100 consumidores foram solicitados a pegar o desenho que preferiam. Foram obtidos os seguintes dados. Os dados confirmam a crença de que um desenho tem a mesma probabilidade de ser escolhido que outro qualquer? Explique.

Desenho	Número de Vezes em que Foi Preferido
1	5
2	15
3	30
4	40
5	10

4.2 EVENTOS E SUAS PROBABILIDADES

Na introdução deste capítulo, utilizamos o termo *evento* de modo muito similar ao usado na linguagem do cotidiano. Depois, na Seção 4.1, introduzimos o conceito de experimento e seus resultados experimentais ou pontos amostrais correspondentes. Pontos amostrais e eventos constituem a base para o estudo das probabilidades. Em conseqüência, precisamos introduzir agora a definição formal de **evento**, uma vez que ele se relaciona aos pontos amostrais. Isso nos dará a base para determinarmos a probabilidade de um evento.

EVENTO

Um evento é um conjunto de pontos amostrais.

Como exemplo, retornemos ao projeto da KP&L e suponhamos que o gerente de projetos esteja interessado na eventualidade de o projeto inteiro ser concluído em dez meses ou menos. Consultando a Tabela 4.3, notamos que seis pontos amostrais – (2, 6), (2, 7), (2, 8), (3, 6), (3, 7) e (4, 6) – apresentam um prazo

de término do projeto de dez meses ou menos. Se considerarmos que C denota a eventualidade de o projeto ser concluído em dez meses ou menos, escrevemos:

$$C = \{(2, 6), (2, 7), (2, 8), (3, 6), (3, 7), (4, 6)\}$$

Considera-se que o evento C ocorra se, desses seis pontos amostrais, *qualquer* um aparecer como resultado experimental.

Dentre outros eventos que poderiam interessar à gerência da KP&L incluem-se os seguintes:

L = a eventualidade de o projeto ser concluído em *menos* de dez meses

M = a eventualidade de o projeto ser concluído em *mais* de dez meses

Usando a informação da Tabela 4.3, notamos que esses eventos consistem nos seguintes pontos amostrais:

$$L = \{(2, 6), (2, 7), (3, 6)\}$$
$$M = \{(3, 8), (4, 7), (4, 8)\}$$

Uma série de eventos adicionais pode ser definida para o projeto da KP&L, mas, em cada caso, o evento deve ser identificado como um conjunto de pontos amostrais do experimento.

Dadas as probabilidades dos pontos amostrais apresentados na Tabela 4.3, podemos usar a seguinte definição para calcular a probabilidade de qualquer evento que a gerência da KP&L possa querer considerar:

PROBABILIDADE DE UM EVENTO

A probabilidade de um evento é igual à soma das probabilidades dos pontos amostrais do evento.

Usando essa definição, calculamos a probabilidade de um evento em particular somando as probabilidades dos pontos amostrais (resultados experimentais) que compõem o evento. Agora podemos calcular a probabilidade de que o projeto demandará dez meses ou menos para ser concluído. Uma vez que esse evento é dado por $C = \{(2, 6), (2, 7), (2, 8), (3, 6), (3, 7), (4, 6)\}$, a probabilidade do evento C, denotada por $P(C)$, será dada por

$$P(C) = P(2, 6) + P(2, 7) + P(2, 8) + P(3, 6) + P(3, 7) + P(4, 6)$$

Referindo-se às probabilidades dos pontos amostrais da Tabela 4.3; temos, portanto,

$$P(C) = 0{,}15 + 0{,}15 + 0{,}05 + 0{,}10 + 0{,}20 + 0{,}05 = 0{,}70$$

Similarmente, desde que a eventualidade de o projeto ser concluído em menos de dez meses seja dada por $L = \{(2, 6), (2, 7), (3, 6)\}$, a probabilidade desse evento é dada por

$$P(L) = P(2, 6) + P(2, 7) + P(3, 6)$$
$$= 0{,}15 + 0{,}15 + 0{,}10 = 0{,}40$$

Finalmente, para a eventualidade de o projeto ser concluído em mais de dez meses, temos $M = \{(3, 8), (4, 7), (4, 8)\}$ e, assim,

$$P(M) = P(3,8) + P(4,7) + P(4,8)$$
$$= 0{,}05 + 0{,}10 + 0{,}15 = 0{,}30$$

Usando esses resultados probabilísticos, agora podemos dizer à gerência da KP&L que há uma probabilidade de 0,70 de o projeto ser concluído em dez meses ou menos, uma probabilidade 0,40 de o projeto ser concluído em menos de dez meses e uma probabilidade de 0,30 de o projeto ser concluído em mais de dez meses. Esse procedimento para o cálculo de probabilidades pode ser repetido para qualquer evento que interesse à gerência da KP&L.

Sempre que podemos identificar todos os pontos amostrais de um experimento e atribuir probabilidades a cada um, temos condições de calcular a probabilidade de determinado evento usando a definição. Entretanto, em muitos experimentos o grande número de pontos amostrais torna a identificação dos pontos amostrais, bem como a determinação de suas respectivas probabilidades, extremamente complicadas, quando não impossíveis. Nas seções restantes deste capítulo apresentaremos algumas relações probabilísticas básicas que podem ser usadas no cálculo da probabilidade de um evento sem a necessidade de se conhecer todas as probabilidades dos pontos amostrais.

NOTAS E COMENTÁRIOS

1. O espaço amostral, S, é um evento. Uma vez que ele contém todos os resultados experimentais, tem a probabilidade 1; ou seja, $P(S) = 1$.
2. Quando se usa o método clássico para atribuir probabilidades, o pressuposto é que os resultados experimentais sejam igualmente prováveis. Nesses casos, a probabilidade de um evento pode ser calculada contando-se o número de resultados experimentais do evento e dividindo-se o resultado pelo número total de resultados experimentais.

Exercícios

Métodos

14. Um experimento tem quatro resultados igualmente prováveis: E_1, E_2, E_3 e E_4.
    a. Qual é a probabilidade de E_2 ocorrer?
    b. Qual é a probabilidade de dois resultados quaisquer ocorrerem (por exemplo, E_1 ou E_3)?
    c. Qual é a probabilidade de três resultados quaisquer ocorrerem (por exemplo, E_1, E_2 ou E_4)?
15. Considere o experimento de escolher uma carta de um baralho de 52 cartas. Cada carta corresponde a um ponto amostral com uma probabilidade de 1/52.
    a. Relacione os pontos amostrais relativos à eventualidade de um ás ser escolhido.
    b. Relacione os pontos amostrais relativos à eventualidade de uma carta com naipe de paus ser escolhida.
    c. Relacione os pontos amostrais relativos à eventualidade de uma das cartas da corte (valete, rainha ou rei) ser escolhida.
    d. Encontre as probabilidades associadas a cada um dos eventos das questões (a), (b) e (c).

AUTOTESTE

16. Considere o evento de lançar um par de dados. Suponha que estejamos interessados na soma dos valores de face mostrados nos dados.
    a. Quantos pontos amostrais são possíveis? (*Dica*: Use a regra de contagem de experimentos em múltiplas etapas.)
    b. Relacione os pontos amostrais.
    c. Qual é a probabilidade de se obter o valor 7?
    d. Qual é a probabilidade de se obter o valor 9 ou um valor maior?
    e. Uma vez que cada lançamento tem seis valores pares possíveis (2, 4, 6, 8, 10 e 12) e somente cinco valores ímpares possíveis (3, 5, 7, 9 e 11) os dados exibirão valores pares com mais freqüência do que valores ímpares. Você concorda com essa afirmação? Explique.
    f. Qual método você usou para atribuir as probabilidades solicitadas?

Aplicações

17. Consulte os pontos amostrais e as probabilidades dos pontos amostrais correspondentes à KP&L indicados nas Tabelas 4.2 e 4.3, respectivamente.
    a. A fase de projeto (etapa 1) estourará o orçamento se demandar quatro meses para ser concluída. Relacione os pontos amostrais relativos à eventualidade de a fase de projeto estourar o orçamento.
    b. Qual é a probabilidade de a fase de projeto estourar o orçamento?
    c. A fase de construção (etapa 2) estourará o orçamento se demandar oito meses para ser concluída. Relacione os pontos amostrais relativos à eventualidade de a etapa de construção estourar o orçamento.
    d. Qual é a probabilidade de a fase de construção estourar o orçamento?
    e. Qual é a probabilidade de ambas as etapas estourarem o orçamento?

AUTOTESTE

18. Suponha que o gerente de um grande complexo de apartamentos forneça as seguintes estimativas de probabilidade subjetivas acerca do número de apartamentos vagos no próximo mês:

Apartamentos Vazios	Probabilidade
0	0,05
1	0,15
2	0,35
3	0,25
4	0,10
5	0,10

Forneça a probabilidade de cada um dos seguintes eventos:

a. Não há apartamentos vazios.
b. Pelo menos quatro apartamentos vazios.
c. Dois ou menos apartamentos vazios.

19. A National Sporting Goods Association realizou uma pesquisa de pessoas com idades a partir de 7 anos sobre a participação em atividades esportivas (*Statistical Abstract of the United States*, 2002). Os dados registrados sobre a população dessa faixa etária indicavam 248,5 milhões de pessoas, sendo 120,9 milhões do sexo masculino e 127,6 milhões do sexo feminino. O número de participantes das cinco principais atividades esportivas é apresentado a seguir:

Atividade	Participantes (milhões)	
	Masculino	Feminino
Andar de bicicleta	22,2	21,0
Acampar	25,6	24,3
Fazer caminhadas	28,7	57,7
Exercitar-se com aparelhos	20,4	24,4
Nadar	26,4	34,4

a. Em relação a pessoas do sexo feminino selecionadas aleatoriamente, estime a probabilidade de participação em cada uma das atividades esportivas.
b. Em relação a pessoas do sexo masculino selecionadas aleatoriamente, estime a probabilidade de participação em cada uma das atividades esportivas.
c. Em relação a uma pessoa selecionada aleatoriamente, estime a probabilidade de ela participar em exercícios de caminhada.
d. Suponha que você acabe de ver passar alguém praticando caminhada. Qual seria a probabilidade de essa pessoa ser uma mulher? Qual seria a probabilidade de essa pessoa ser um homem?

20. A revista *Fortune* publica uma edição anual que contém informações sobre as empresas do grupo *Fortune* 500. Os dados a seguir apresentam os seis estados que contam com o maior número de empresas do grupo *Fortune* 500, bem como o número de empresas cuja sede se encontra nesses estados (*Fortune*, 17 de abril de 2000).

Estado	Número de Empresas
Nova York	56
Califórnia	53
Texas	43
Illinois	37
Ohio	28
Pennsylvania	28

Suponha que uma empresa do grupo *Fortune* 500 seja escolhida para responder a um questionário de delineamento (*follow-up*). Quais são as probabilidades dos seguintes eventos?

a. Seja N a eventualidade de a empresa ter sede em Nova York. Encontre $P(N)$.
b. Seja T a eventualidade de a empresa ter sede no Texas. Encontre $P(T)$.
c. Seja B a eventualidade de a empresa ter sede em um desses seis estados. Encontre $P(B)$.

21. A população norte-americana (Estados Unidos), distribuída por faixa etária, é a seguinte (*The World Almanac 2004*). Os dados estão expressos em milhões de pessoas.

Idade	Número
19 anos ou menos	80,5
20 a 24	19,0
25 a 34	39,9
35 a 44	45,2
45 a 54	37,7
55 a 64	24,3
65 anos ou mais	35,0

Suponha que uma pessoa seja escolhida aleatoriamente dessa população.

a. Qual é a probabilidade de a pessoa ter de 20 a 24 anos?
b. Qual é a probabilidade de a pessoa ter de 20 a 34 anos?
c. Qual é a probabilidade de a pessoa ter acima de 45 anos?

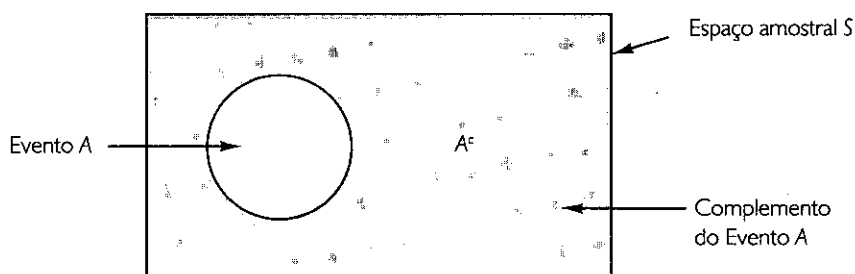
4.3 ALGUMAS RELAÇÕES BÁSICAS DE PROBABILIDADE

Complemento de um Evento

Dado um evento A, o **complemento de** A é definido como o evento que consiste em todos os pontos amostrais que *não* estão em A. O complemento de A é denotado por A^c. A Figura 4.4 representa um diagrama conhecido como **diagrama de Venn**, que ilustra o conceito de complemento. A área retangular representa o espaço amostral do experimento e, como tal, contém todos os pontos amostrais possíveis. O círculo representa o evento A e contém somente os pontos amostrais que pertencem a A. A região sombreada do retângulo contém todos os pontos amostrais que não estão no evento A e, por definição, é o complemento de A.

Em qualquer aplicação de probabilidade, ou o evento A ou o seu complemento A^c devem ocorrer. Portanto, temos

$$P(A) + P(A^c) = 1$$

Figura 4.4 A área sombreada é o complemento do evento A

Resolvendo $P(A)$, obtemos o resultado seguinte.

COMO CALCULAR A PROBABILIDADE USANDO O COMPLEMENTO

$$P(A) = 1 - P(A^c) \qquad \textbf{(4.5)}$$

A Equação 4.5 mostra que a probabilidade de um evento A pode ser facilmente calculada se a probabilidade de seu complemento, $P(A^c)$, for conhecida.

Como exemplo, considere o caso de um gerente de vendas que, após revisar os relatórios, afirma que 80% dos contatos com novos clientes não resultaram em vendas. Se considerarmos A a eventualidade de ocorrer uma venda e A^c a eventualidade de não ocorrer nenhuma venda, o gerente está afirmando que $P(A^c) = 0{,}80$. Usando a Equação 4.5, vemos que

$$P(A) = 1 - P(A^c) = 1 - 0{,}80 = 0{,}20$$

Podemos concluir que o contato com novos clientes tem 0,20 de probabilidade de resultar em uma venda.

Em outro exemplo, um agente de compras afirma que probabilidade de o fornecedor enviar uma remessa isenta de peças defeituosas é 0,90. Usando o complemento, podemos concluir que probabilidade de a remessa não conter peças defeituosas é $1 - 0{,}90 = 0{,}10$.

Lei da Adição

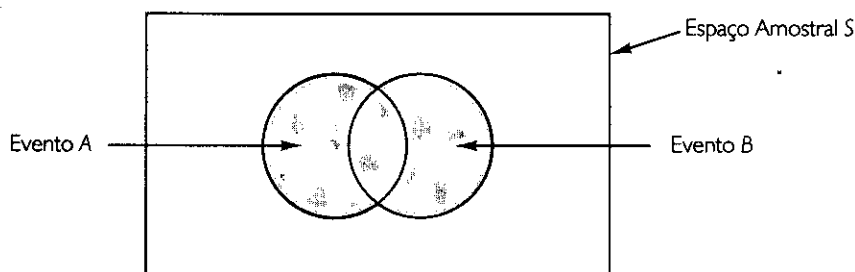
A lei da adição é útil quando estamos interessados em saber qual é a probabilidade de pelo menos um de dois eventos ocorrer. Ou seja, com os eventos A e B estamos interessados em saber qual é a probabilidade de ocorrência do evento A ou do evento B, ou de ambos.

Antes de apresentarmos a lei da adição, precisamos discutir dois conceitos relacionados à combinação de eventos: a *união* de eventos e a *interseção de eventos*. Dados dois eventos A e B, a **união de A e B** é definida da seguinte maneira:

UNIÃO DE DOIS EVENTOS

A *união* de A e B é o evento que contém *todos* os pontos amostrais que pertencem a A *ou* B, *ou a ambos*. A união é denotada por $A \cup B$.

O diagrama de Venn da Figura 4.5 retrata a união dos eventos A e B. Observe que os dois círculos contêm todos os pontos amostrais do evento A, bem como os pontos amostrais do evento B.

Figura 4.5 A área sombreada é a união dos eventos A e B

O fato de os círculos se sobreporem indica que alguns pontos amostrais estão contidos tanto em A como em B.

A definição da **interseção de A e B** é a seguinte:

INTERSEÇÃO DE DOIS EVENTOS

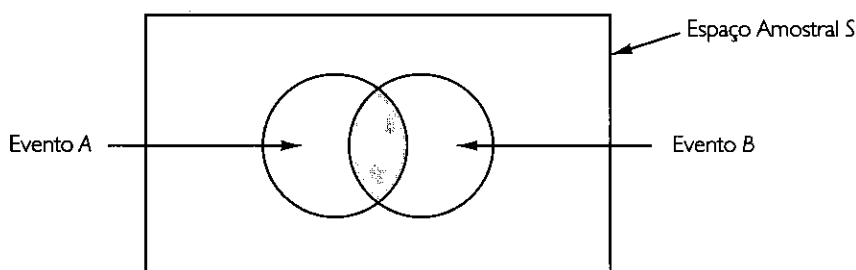
Dados dois eventos A e B, a *interseção* de A e B é o evento que contém os pontos amostrais que pertencem *tanto* a A *como* a B. A interseção é denotada por $A \cap B$.

O diagrama de Venn que retrata a interseção dos eventos A e B é mostrado na Figura 4.6. A área em que os dois círculos se sobrepõem é a interseção; ela contém os pontos amostrais que estão tanto em A como em B.

Vamos prosseguir agora com a discussão da lei da adição. A **lei da adição** constitui uma maneira de calcular a probabilidade de o evento A ou o evento B, ou ambos, ocorrerem. Em outras palavras, a lei da adição é usada para calcular a probabilidade da união de dois eventos. A lei da adição é escrita da seguinte maneira:

LEI DA ADIÇÃO

$$P(A \cap B) = P(A) + P(B) - P(A \cap B) \qquad \textbf{(4.6)}$$

Figura 4.6 A área sombreada é a interseção dos eventos A e B

Para entender a lei da adição intuitivamente, observe que os dois primeiros termos da lei da adição, $P(A) + P(B)$, contabilizam todos os pontos amostrais de $A \cup B$. Entretanto, desde que os pontos amostrais na interseção $A \cap B$ estão tanto em A como em B, quando calculamos $P(A) + P(B)$, estamos efetivamente contando cada um dos pontos amostrais em $A \cap B$ duas vezes. Corrigimos essa contagem em dobro ao subtrair $P(A \cap B)$.

Como exemplo da aplicação da lei da adição, consideremos o caso de uma pequena planta de montagem com 50 empregados. Espera-se que cada funcionário conclua suas obrigações no prazo e que as desempenhe de tal maneira que o produto montado seja aprovado na inspeção final. Ocasionalmente, algum funcionário deixa de cumprir os padrões de desempenho, concluindo o trabalho tardiamente ou montando produtos com defeito. Ao final de um período de avaliação do desempenho, o gerente de produção descobriu que cinco dos 50 funcionários concluíam o trabalho atrasados e seis dos 50 montavam um produto com defeito e dois dos 50 funcionários tanto concluíam o trabalho tardiamente como montando produtos com defeitos.

Admitamos que

L = a eventualidade de o trabalho ser concluído atrasado

D = a eventualidade de o produto montado apresentar defeito

A informação sobre a freqüência relativa nos leva às seguintes probabilidades.

$$P(L) = \frac{5}{50} = 0{,}10$$

$$P(D) = \frac{6}{50} = 0{,}12$$

$$P(L \cap D) = \frac{2}{50} = 0{,}04$$

Depois de revisar os dados de desempenho, o gerente de produção decidiu atribuir avaliações de desempenho a qualquer empregado cujo trabalho fosse concluído atrasado ou apresentando defeitos; desse modo, o evento de interesse é $L \cup D$. Qual é a probabilidade de o gerente de produção atribuir uma avaliação ruim a um funcionário?

Observe que a questão probabilística se refere à união de dois eventos. Especificamente, queremos conhecer

$$P(L \cup D) = P(L) + P(D) - P(L \cap D)$$

Conhecendo os valores das três probabilidades expressas no segundo membro dessa equação, podemos escrever

$$P(L \cup D) = 0{,}10 + 0{,}12 - 0{,}04 = 0{,}18$$

Esse cálculo nos informa que há 0,18 de probabilidade de que um funcionário escolhido aleatoriamente receba uma classificação de desempenho ruim.

Como outro exemplo da lei da adição, considere um estudo realizado recentemente pelo gerente de pessoal de uma grande empresa de software de computador. O estudo mostrou que 30% dos funcionários que saíram da firma no intervalo de dois anos o fizeram porque estavam insatisfeitos com seus salários, 20% saíram porque estavam insatisfeitos com suas atribuições de trabalho e 12% dos ex-funcionários indicaram insatisfação tanto com o salário como com suas atribuições de trabalho. Qual é a probabilidade de um funcionário que sair dentro de dois anos vir a fazê-lo em virtude da insatisfação com o salário, insatisfação com a atribuição de trabalho, ou ambos?

Admitamos que

S = a eventualidade de o empregado sair em razão do salário

W = a eventualidade de o empregado sair em decorrência da atribuição de trabalho

Temos $P(S) = 0{,}30$, $P(W) = 0{,}20$ e $P(S \cap W) = 0{,}12$. Usando a Equação 4.6, a lei da adição, temos

$$P(S \cup W) = P(S) + P(W) - P(S \cap W) = 0{,}30 + 0{,}20 - 0{,}12 = 0{,}38$$

Descobrimos que há uma probabilidade de 0,38 de que um funcionário saia da empresa por motivos de salário ou de atribuição funcional.

Antes de concluirmos nossa discussão da lei de adição, vamos considerar um caso especial que se apresenta para **eventos mutuamente exclusivos.**

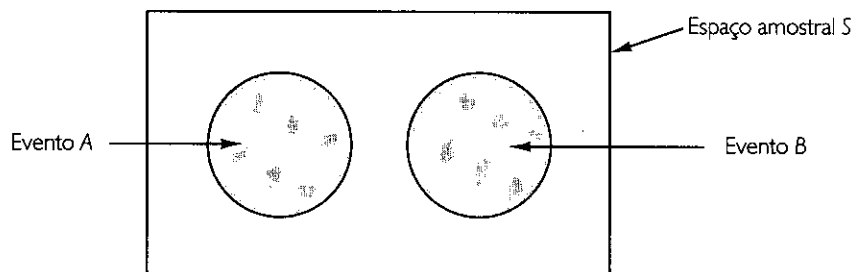
EVENTOS MUTUAMENTE EXCLUSIVOS

Dois eventos são considerados mutuamente exclusivos se eles não tiverem nenhum ponto amostral em comum.

Os eventos A e B são mutuamente exclusivos se, quando um evento ocorre, o outro não pode ocorrer. Assim, um requisito para A e B serem mutuamente exclusivos é que sua interseção não deve conter nenhum ponto amostral. O diagrama de Venn que descreve dois eventos A e B mutuamente exclusivos é apresentado na Figura 4.7. Nesse caso, $P(A \cap B) = 0$ e a lei da adição pode ser escrita da seguinte maneira:

LEI DA ADIÇÃO PARA EVENTOS MUTUAMENTE EXCLUSIVOS

$$P(A \cup B) = P(A) + P(B)$$

Figura 4.7 Eventos mutuamente exclusivos

Exercícios

Métodos

22. Suponha que temos um espaço amostral com cinco resultados experimentais igualmente prováveis: E_1, E_2, E_3, E_4, E_5. Admitamos que

$$A = \{E_1, E_2\}$$
$$B = \{E_3, E_4\}$$
$$C = \{E_2, E_3, E_5\}$$

a. Encontre $P(A)$, $P(B)$ e $P(C)$.
b. Encontre $P(A \cup B)$. A e B são mutuamente exclusivos?
c. Encontre A^c, C^c, $P(A^c)$ e $P(C^c)$.
d. Encontre $A \cup B^c$ e $P(A \cup B^c)$.
e. Encontre $P(B \cup C)$.

AUTOTESTE

23. Suponha que temos um espaço amostral $S = \{E_1, E_2, E_3, E_4, E_5, E_6, E_7\}$, em que $E_1, E_2, \ldots E_7$ denotam os pontos amostrais. Aplicam-se as seguintes atribuições de probabilidade: $P(E_1) = 0{,}05$, $P(E_2) = 0{,}20$, $P(E_3) = 0{,}20$, $P(E_4) = 0{,}25$, $P(E_5) = 0{,}15$, $P(E_6) = 0{,}10$ e $P(E_7) = 0{,}05$. Admitamos que

$$A = \{E_1, E_4, E_6\}$$
$$B = \{E_2, E_4, E_7\}$$
$$C = \{E_2, E_3, E_5, E_7\}$$

a. Encontre $P(A)$, $P(B)$ e $P(C)$.
b. Ache $A \cup B$ e $P(A \cup B)$.
c. Encontre $A \cap B$ e $P(A \cap B)$.
d. Os eventos A e C são mutuamente exclusivos?
e. Encontre B^c e $P(B^c)$.

Aplicações

24. A Clarkson University fez uma pesquisa de seus ex-formandos para conhecer melhor o que eles pensam a respeito da universidade. Uma parte da pesquisa pedia que os consultados indicassem se a experiência que haviam tido na Clarkson ficara aquém das expectativas, se atingira as expectativas ou se superara as expectativas. O resultado mostrou que 4% dos consultados nada responderam, 26% disseram que suas experiências ficaram aquém das expectativas e 65% dos consultados disseram que suas experiências atingiram as expectativas (*Clarkson Magazine*, verão de 2001).
    a. Se escolhermos um ex-aluno aleatoriamente, qual é a probabilidade de ele afirmar que sua experiência superou as expectativas?
    b. Se escolhermos um ex-aluno aleatoriamente, qual é a probabilidade de ele afirmar que suas expectativas foram atingidas ou superadas.

25. Dados divulgados sobre os 30 maiores fundos de ações apresentaram a rentabilidade percentual para aplicações de um ano e de cinco anos, respectivamente, correspondentes ao período com vencimento em 31 de março de 2000 (*The Wall Street Journal*, 10 de abril de 2000). Suponha que consideremos elevada uma rentabilidade superior a 2% para aplicações de um ano e que consideremos também elevada uma rentabilidade acima de 44% para aplicações de cinco anos. Metade dos fundos teve rentabilidade acima de 2% para aplicações de um ano, 12% dos fundos tiveram rentabilidade acima de 44% para aplicações de cinco anos, e seis dos fundos tanto tiveram rentabilidade acima de 2% para aplicações de um ano como rentabilidade acima de 44% para aplicações de cinco anos.
    a. Encontre a probabilidade de um fundo ter uma rentabilidade elevada para aplicações de um ano, a probabilidade de um fundo ter uma rentabilidade elevada para aplicações de cinco anos, e a probabilidade de um fundo ter tanto uma rentabilidade elevada para aplicações de um ano como uma rentabilidade elevada para aplicações de cinco anos.
    b. Qual é a probabilidade de um fundo ter obtido uma rentabilidade elevada para aplicações de um ano, uma rentabilidade elevada para aplicações de cinco anos, ou ambos?
    c. Qual é a probabilidade de um fundo não ter obtido uma rentabilidade elevada tanto para aplicações de um ano como para as de cinco anos?

26. Dados divulgados sobre os 30 maiores fundos de ações e de investimentos diversificados apresentaram a rentabilidade percentual para aplicações de um ano e de cinco anos, respectivamente, correspondentes ao período com vencimento em 31 de março de 2000 (*The Wall Street Journal*, 10 de abril de 2000). Suponha que consideremos elevada uma rentabilidade superior a 50%, para aplicações de um ano e que consideremos também elevada uma rentabilidade acima de 300%, para aplicações de cinco anos. Nove dos fundos tiveram rentabilidade acima de 50% para aplicações de um ano, sete dos fundos tiveram rentabilidade acima de 300%, para aplicações de cinco anos, e cinco dos fundos tanto tiveram rentabilidade acima de 50% para aplicações de um ano, como rentabilidade acima de 300% para aplicações de cinco anos.
    a. Qual é a probabilidade de haver uma rentabilidade elevada para aplicações de um ano, e qual é a probabilidade de rentabilidade elevada para aplicações de cinco anos?
    b. Qual é a probabilidade de rentabilidade elevada tanto para aplicações de um ano como para aplicações de cinco anos?
    c. Qual é a probabilidade de não haver rentabilidade elevada para aplicações de um ano nem para aplicações de cinco anos?

27. Uma pesquisa de opinião realizada na pré-temporada do campeonato de futebol norte-americano da NCAA pediu aos consultados para responderem à seguinte pergunta: "A Conferência de Atletismo Big Ten ou a Pac-10 terá um time no jogo de decisão do campeonato nacional deste ano, a Rose Bowl?" Dos 13.429 consultados, 2.961 disseram que a Big 10 teria, 4.494 disseram que a Pac-10 teria e 6.823 disseram que nem a Big Ten nem a Pac-10 teriam um time na Rose Bowl (www.yahoo.com, 30 de agosto de 2001).

a. Qual é a probabilidade de o consultado ter respondido que nem a Big Ten nem a Pac-10 terá um time na Rose Bowl?
b. Qual é a probabilidade de o consultado ter respondido que ou a Big Ten ou a Pac-10 terá um time na Rose Bowl?
c. Encontre a probabilidade de o consultado ter respondido que tanto a Big Ten como a Pac-10 terão um time na Rose Bowl?

AUTOTESTE

28. Uma pesquisa de assinantes de revista mostrou que 45,8% alugaram um carro nos últimos 12 meses por razões comerciais, 54% alugaram um carro durante os últimos 12 meses por razões pessoais e 30% alugaram um carro nos últimos 12 meses tanto por razões comerciais como por razões pessoais.
   a. Qual é a probabilidade de um assinante ter alugado um carro durante os últimos 12 meses por razões comerciais ou pessoais?
   b. Qual é a probabilidade de um assinante não ter alugado um carro durante os últimos 12 meses por razões comerciais ou por razões pessoais?
29. Estudantes que concluem a fase *sênior*[1] do curso colegial com ótimo desempenho candidatam-se aos cursos universitários mais seletivos em um número cada vez maior a cada ano. Uma vez que o número de vagas permanece relativamente estável, algumas universidades recusam um número maior dos *early applicants*[2]. A Universidade da Pensilvânia recebeu 2.851 inscrições de *early applicants*. Desse grupo, admitiu 1.033 estudantes, recusou 854 imediatamente e protelou 964 para o *pool* de admissões normais. A universidade admitiu cerca de 18% dos candidatos do *pool* de admissões normais considerando um tamanho total de classes (número de admissões de *early apllicants* mais as admissões normais) de 2.375 estudantes (*USA Today*, 24 de janeiro de 2001). Vamos considerar que E, R e D representam a eventualidade (eventos) de um estudante que se candidata à *early admission* ser admitido, recusado imediatamente ou protelado para o *pool* de admissões normais; e que A representa a eventualidade (evento) de um estudante ser admitido no conjunto de admissões normais.
   a. Use os dados para estimar $P(E)$, $P(R)$ e $P(D)$.
   b. Os eventos E e D são mutuamente exclusivos? Encontre $P(E \cap D)$.
   c. Em relação aos 2.375 estudantes admitidos na Universidade da Pensilvânia, qual é a probabilidade de um estudante escolhido aleatoriamente ter sido aceito para *early admission*?
   d. Suponha que um estudante se inscreva na Universidade da Pensilvânia para *early admission*. Qual é a probabilidade de o estudante ser admitido para *early admission* ou ser aceito para admissão no *pool* de admissões normais?

4.4 PROBABILIDADE CONDICIONAL

Freqüentemente, a probabilidade de um evento é influenciada pelo fato de um evento relacionado já ter ocorrido ou não. Suponha que temos um evento A com a probabilidade $P(A)$. Se obtivermos uma nova informação e soubermos que um evento relacionado, denotado por B, já ocorreu, quereremos tirar proveito dessa informação calculando uma nova probabilidade para o evento A.

Essa nova probabilidade do evento A denomina-se **probabilidade condicional** e é escrita como $P(A \mid B)$. Usamos a notação $\mid$ para indicar que estamos considerando a probabilidade do evento A *dada* a condição de o evento B ter ocorrido. Portanto, a notação $P(A \mid B)$ é lida da seguinte maneira: "a probabilidade de A, dado B."

Como ilustração da aplicação da probabilidade condicional, considere a situação do *status* de promoção de oficiais masculinos e femininos de um grande departamento de polícia metropolitana no leste dos Estados Unidos. A força policial consiste em 1.200 oficiais, sendo 960 homens e 240 mulheres. Nos últimos dois anos, 324 oficiais da força policial receberam promoções. A estrutura específica de promoções para oficiais masculinos e femininos é apresentada na Tabela 4.4.

---

[1] NT: Após seis anos de *elementary school*, na qual o aluno aprende as matérias básicas, ele segue para o curso secundário, ou *high school*, que consiste na *junior high school*, com duração de três anos, e depois a *senior high school*, que oferece o último ano da educação colegial.

[2] NT: Um *early applicant* é definido como o estudante que deseja ingressar na universidade após a conclusão da etapa *júnior* do curso colegial (*high school*).

Depois de rever o registro de promoções, uma comissão de oficiais femininas fez uma acusação formal de discriminação baseando-se no fato de que 288 oficiais masculinos receberam promoções e somente 36 oficiais femininas foram promovidas. A administração da polícia argüiu que o número relativamente baixo de promoções para as oficiais femininas se deveu não à discriminação, mas ao fato de relativamente poucas mulheres serem integrantes da força policial. Vamos mostrar como a probabilidade condicional poderia ser usada para analisar a acusação de discriminação.

Se admitirmos que

H = o evento de um oficial ser homem
M = o evento de um oficial ser mulher
A = o evento de um oficial ser promovido
A^C = o evento de um oficial não ser promovido

Dividir os valores de dados da Tabela 4.4 pelo total de 1.200 oficiais nos possibilita sintetizar a informação disponível com os seguintes valores probabilísticos:

$P(H \cap A) = 288/1.200 = 0{,}24$ = probabilidade de um oficial escolhido aleatoriamente ser um homem *e* ser promovido.
$P(H \cap A^c) = 672/1.200 = 0{,}56$ = probabilidade de um oficial escolhido aleatoriamente ser um homem *e* não ser promovido.
$P(M \cap A) = 36/1.200 = 0{,}03$ = probabilidade de um oficial escolhido aleatoriamente ser uma mulher *e* ser promovida.
$P(M \cap A^c) = 204/1.200 = 0{,}17$ = probabilidade de um oficial escolhido aleatoriamente ser uma mulher *e* não ser promovida.

Uma vez que cada um desses valores dá a probabilidade da interseção de dois eventos, as probabilidades são chamadas **probabilidades associadas.** A Tabela 4.5, que apresenta um resumo das informações probabilísticas referentes à situação das promoções dos oficiais do departamento de polícia, é denominada *tabela de probabilidade associada.*

Os valores indicados nas margens da tabela de probabilidade associada fornecem as probabilidades de cada evento separadamente. Ou seja, $P(H) = 0{,}80$, $P(M) = 0{,}20$, $P(A) = 0{,}27$ e $P(A^c) = 0{,}73$. Essas probabilidades se denominam **probabilidades marginais** em virtude de sua localização nas margens da tabela de probabilidade associada.

Tabela 4.4 *Status* de promoção dos oficiais de polícia nos dois últimos anos

	Homens	Mulheres	Total
Promovidos	288	36	324
Não promovidos	672	204	876
Total	960	240	1.200

Tabela 4.5 Tabela de probabilidade associada das promoções

As probabilidades associadas aparecem no corpo da tabela.

	Homens (H)	Mulheres (M)	Total
Promovidos (A)	0,24	0,03	0,27
Não promovidos (A^C)	0,56	0,17	0,73
Total	0,80	0,20	1,00

As probabilidades marginais aparecem nas margens da tabela.

Notamos que as probabilidades associadas são encontradas somando-se as probabilidades associadas que se encontram na linha ou coluna correspondentes da tabela de probabilidade associada. Por exemplo, a probabilidade marginal de alguém ser promovido é $P(A) = P(H \cap A) + P(M \cap A) = 0{,}24 + 0{,}03 = 0{,}27$. Das probabilidades marginais, vemos que 80% da força policial são homens, 20% da força são mulheres, 27% de todos os oficiais receberam promoções e 73% não foram promovidos.

Vamos iniciar a análise da probabilidade condicional calculando a probabilidade de um oficial ser promovido dado que o oficial seja um homem. Na notação de probabilidade condicional, tentamos determinar $P(A \mid H)$. Para calcular $P(A \mid H)$, primeiramente precisamos entender que essa notação significa simplesmente que estamos considerando a probabilidade do evento A (promoção), visto que sabemos da existência da condição designada como evento H (o oficial ser um homem). Assim, $P(A \mid H)$ nos diz que agora estamos interessados somente no *status* de promoção dos 960 oficiais do sexo masculino. Uma vez que 288 dos 960 oficiais do sexo masculino receberam promoções, a probabilidade de haver uma promoção dado que o oficial seja um homem é 288/960 = 0,30. Em outras palavras, dado que um oficial seja um homem, ele teve 30% de chance de receber uma promoção no decorrer dos últimos dois anos.

Esse procedimento foi fácil de aplicar porque os valores apresentados na Tabela 4.4 mostram o número de oficiais de cada categoria. Queremos demonstrar agora como se pode calcular diretamente probabilidades condicionais como $P(A \mid H)$, a partir das probabilidades de eventos, em vez dos dados de freqüência da Tabela 4.4.

Mostramos que $P(A \mid H) = 288/960 = 0{,}30$. Vamos dividir agora tanto o numerador como o denominador dessa fração por 1.200, que é o número total de oficiais integrantes do estudo.

$$P(A \mid H) = \frac{288}{960} = \frac{288/1.200}{960/1.200} = \frac{0{,}24}{0{,}80} = 0{,}30$$

Notamos agora que a probabilidade condicional $P(A \mid H)$ pode ser calculada como 0,24/0,80. Consulte a tabela de probabilidade associada (Tabela 4.5). Observe, em especial, que 0,24 é a probabilidade associada de A e H; ou seja, $P(A \cap H) = 0{,}24$. Note também que 0,80 é a probabilidade marginal de um oficial aleatoriamente selecionado ser um homem; ou seja, $P(H) = 0{,}80$. Desse modo, a probabilidade condicional $P(A \mid H)$ pode ser calculada como a razão da probabilidade associada $P(A \cap H)$ pela probabilidade marginal $P(H)$.

$$P(A \mid H) = \frac{P(A \cap H)}{P(H)} = \frac{0{,}24}{0{,}80} = 0{,}30$$

O fato de as probabilidades condicionais poderem ser calculadas como a razão de uma probabilidade associada pela probabilidade marginal nos fornece a seguinte fórmula geral para efetuarmos cálculos da probabilidade condicional de dois eventos A e B.

PROBABILIDADE CONDICIONAL

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)} \tag{4.7}$$

ou

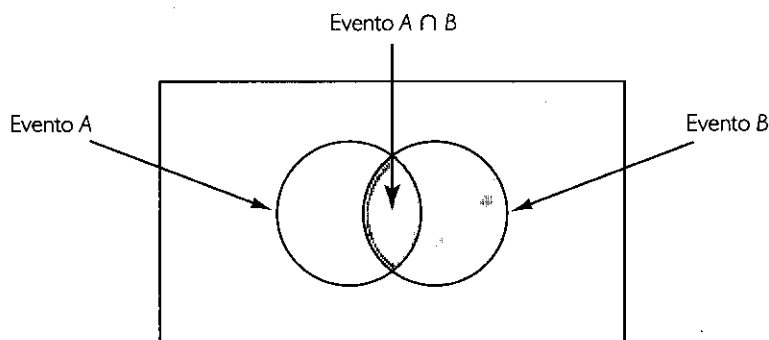
$$P(B \mid A) = \frac{P(A \cap B)}{P(A)} \tag{4.8}$$

O diagrama de Venn da Figura 4.8 é útil para obtermos um entendimento intuitivo da probabilidade condicional. O círculo à direita mostra que ocorreu o evento B; a parte do círculo que se sobrepõe ao evento A denota o evento $(A \cap B)$. Sabemos que, desde que o evento B ocorreu, a única maneira pela qual também podemos observar o evento A é pela ocorrência do evento $(A \cap B)$. Assim, a razão $P(A \cap B)/P(B)$ nos fornece a probabilidade condicional de que observaremos o evento A dado que o evento B já ocorreu.

Retornemos à questão da discriminação contra oficiais do sexo feminino. A probabilidade marginal apresentada na linha 1 da Tabela 4.5 nos mostra que a probabilidade de promoção de um oficial é $P(A) = 0{,}27$ (independentemente de o oficial ser homem ou mulher). Entretanto, a questão crucial no caso da discriminação envolve as duas probabilidades condicionais $P(A \mid H)$ e $P(A \mid M)$. Ou seja, qual é a pro-

babilidade de promoção *dado* que o oficial seja um homem, e qual é a probabilidade de promoção *dado* que o oficial seja uma mulher? Se essas duas probabilidades forem iguais, não há base para o argumento de discriminação porque as chances de promoção são as mesmas para oficiais do sexo masculino e do sexo feminino. No entanto, a diferença nas duas probabilidades condicionais sustentará a posição de que os oficiais masculinos e femininos são tratados diferentemente nas decisões de promoção.

Já determinamos que $P(A \mid H) = 0{,}30$. Vamos usar agora os valores de probabilidade da Tabela 4.5 e a relação básica da probabilidade condicional apresentada na Equação 4.7 para calcular a probabilidade de um oficial ser promovido, dado que o oficial seja uma mulher; ou seja, $P(A \mid M)$.

Figura 4.8 Probabilidade condicional $P(A \mid B) = P(A \cap B)/P(B)$

Usando a Equação 4.7, com M substituindo H, obtemos:

$$P(A \mid M) = \frac{P(A \cap M)}{P(M)} = \frac{0{,}03}{0{,}20} = 0{,}15$$

Que conclusão você tira? A probabilidade de haver uma promoção, dado que o oficial seja homem é de 0,30, duas vezes a probabilidade de 0,15 de promoção, dado que o oficial seja uma mulher. Não obstante o uso da probabilidade condicional não provar por si mesmo que exista discriminação nesse caso, os valores da probabilidade condicional sustentam o argumento apresentado pelas oficiais.

Eventos Independentes

Na ilustração anterior, $P(A) = 0{,}27$, $P(A \mid H) = 0{,}30$ e $P(A \mid M) = 0{,}15$. Notamos que a probabilidade de uma promoção (evento A) é afetada ou influenciada pelo fato de o oficial ser um homem ou uma mulher. Especialmente, desde que $P(A \mid H) \neq P(A)$, poderíamos dizer que os eventos A e H são eventos dependentes, isto é, a probabilidade do evento A (promoção) é alterada ou afetada pelo fato de se saber que o evento H (o oficial é um homem) existe. Analogamente, com $P(A \mid M) \neq P(A)$, poderíamos dizer que os eventos A e M são *eventos dependentes*. Entretanto, se há probabilidade de o evento A não se alterar em função da existência do evento H – ou seja, $P(A \mid H) = P(A)$ –, diríamos que os eventos A e H são **eventos independentes**. Essa situação leva à seguinte definição de independência de dois eventos:

EVENTOS INDEPENDENTES

Dois eventos A e B são independentes se

$$P(A \mid B) = P(A) \quad \textbf{(4.9)}$$

ou

$$P(B \mid A) = P(B) \quad \textbf{(4.10)}$$

Caso contrário, os eventos são dependentes.

Lei da Multiplicação

Enquanto a lei da adição é usada para calcular a probabilidade de uma união de dois eventos, a lei da multiplicação é usada para calcular a probabilidade de uma interseção de dois eventos. A lei da multiplicação

baseia-se na definição da probabilidade condicional. Usando as Equações 4.7 e 4.8 e resolvendo $P(A \cap B)$, obtemos a **lei da multiplicação.**

LEI DA MULTIPLICAÇÃO

$$P(A \cap B) = P(B)P(A \mid B) \qquad \textbf{(4.11)}$$

ou

$$P(A \cap B) = P(A)P(B \mid A) \qquad \textbf{(4.12)}$$

Para ilustrarmos o uso da lei da multiplicação, considere o departamento de circulação de um jornal, sabendo-se que 84% das famílias de determinado bairro assinam a edição diária do jornal. Se admitirmos que D denota o evento de uma família assinar a edição diária, $P(D) = 0,84$. Além disso, sabe-se que a probabilidade de uma família que já tem uma assinatura da edição diária também assinar a edição de domingo (evento S) é 0,75; ou seja, $P(S \mid D) = 0,75$.

Qual é a probabilidade de uma família assinar tanto a edição diária como a edição de domingo do jornal? Usando a lei da multiplicação, calculamos a $P(S \cap D)$ desejada como

$$P(S \cap D) = P(D)P(S \mid D) = 0,84(0,75) = 0,63$$

Sabemos agora que 63% das famílias assinam tanto a edição diária quanto a edição dominical.

Antes de concluirmos esta seção, consideremos o caso especial da lei da multiplicação em que os eventos envolvidos são independentes. Lembre-se de que A e B são eventos independentes quando quer que $P(A \mid B) = P(A)$ ou $P(B \mid A) = P(B)$. Portanto, usando as Equações 4.11 e 4.12 para o caso especial dos eventos independentes, obtemos a seguinte lei da multiplicação.

LEI DA MULTIPLICAÇÃO PARA EVENTOS INDEPENDENTES

$$P(A \cap B) = P(A)P(B) \qquad \textbf{(4.13)}$$

Para calcular a probabilidade da interseção de dois eventos independentes, simplesmente multiplicamos as probabilidades correspondentes. Note que a lei da multiplicação para eventos independentes constitui outra maneira de determinarmos se A e B são independentes. Ou seja, se $P(A \cap B) = P(A)P(B)$, então A e B são independentes; se $P(A \cap B) \neq P(A)P(B)$, então A e B são dependentes.

Como uma aplicação da lei da multiplicação para eventos independentes, considere a situação de um gerente de posto de gasolina que sabe, por experiência, que 80% dos clientes usam cartões de crédito ao comprar gasolina. Qual é a probabilidade de os dois próximos clientes que compram gasolina usarem, cada um, um cartão de crédito? Se admitirmos que

A = o evento de o primeiro cliente usar um cartão de crédito
B = o evento de o segundo cliente usar um cartão de crédito

então o evento que os interessa é $A \cap B$. Sem contarmos com nenhuma outra informação, podemos racionalmente supor que A e B são eventos independentes. Desse modo,

$$P(A \cap B) = P(A)P(B) = (0,80)(0,80) = 0,64$$

Para resumir esta seção, observamos que nosso interesse na probabilidade condicional é motivado pelo fato de os eventos freqüentemente serem relacionados. Nesses casos, dizemos que os eventos são dependentes e as fórmulas apresentadas nas Equações 4.7 e 4.8 devem ser usadas para calcular as probabilidades do evento. Se dois eventos não estão relacionados, eles são independentes; nesse caso, não ocorrem nem a probabilidade do evento nem o outro evento.

NOTAS E COMENTÁRIOS

Não confunda a noção de eventos mutuamente exclusivos com a dos eventos independentes. Dois eventos com probabilidades diferentes de zero não podem ser tanto mutuamente exclusivos como independentes. Quando se sabe que um evento mutuamente exclusivo ocorre, o outro não pode ocorrer; assim, a probabilidade de o outro evento ocorrer é reduzida a zero. Portanto, eles são dependentes.

Exercícios

Métodos

AUTOTESTE

30. Suponha que temos dois eventos, A e B, sendo $P(A) = 0{,}50$, $P(B) = 0{,}60$ e $P(A \cap B) = 0{,}40$.
    a. Encontre $P(A \mid B)$.
    b. Encontre $P(B \mid A)$.
    c. A e B são independentes? Por quê?
31. Suponha que temos dois eventos, A e B, que são mutuamente exclusivos. Suponha, além disso, que conhecemos $P(A) = 0{,}30$ e $P(B) = 0{,}40$.
    a. Qual é $P(A \cap B)$?
    b. Qual é $P(A \mid B)$?
    c. Um estudante de estatística argumenta que os conceitos de eventos mutuamente exclusivos e eventos independentes são, na verdade, o mesmo, e que se os eventos são mutuamente exclusivos eles devem ser independentes. Você concorda com essa afirmação? Use a informação de probabilidade nesse problema para justificar sua resposta.
    d. Qual conclusão geral você tiraria a respeito dos eventos mutuamente exclusivos e dos eventos independentes em função dos resultados desse problema?

Aplicações

32. Em razão do aumento dos custos dos seguros-saúde, 43 milhões de pessoas nos Estados Unidos não têm seguro-saúde (*Time*, 1º de dezembro de 2003). Dados amostrais representativos da cobertura de seguro-saúde em nível nacional para pessoas com idades a partir dos 18 anos são apresentados a seguir:

		Seguro-Saúde	
		Sim	Não
Idade	18 a 34	750	170
	35 ou mais	950	130

    a. Desenvolva uma tabela de probabilidade associada e use-a para responder às questões restantes.
    b. Que as probabilidades marginais lhe informam a respeito da idade da população norte-americana?
    c. Qual é a probabilidade de um indivíduo escolhido aleatoriamente ter cobertura de seguro-saúde?
    d. Se o indivíduo tiver entre 18 e 34 anos, qual é a probabilidade de ele não ter cobertura de seguro-saúde?
    e. Se o indivíduo tiver mais de 35 anos, qual é a probabilidade de ele não ter cobertura de seguro-saúde?
    f. Se o indivíduo não tiver seguro-saúde, qual é a probabilidade de ele estar na faixa etária de 18 a 34 anos?
    g. O que a informação de probabilidade lhe diz sobre a cobertura de seguro-saúde nos Estados Unidos?
33. Em uma pesquisa de estudantes de MBA foram obtidos os seguintes dados a respeito da principal razão pela qual os "estudantes" haviam escolhido a escola na qual se matricularam.

AUTOTESTE

		Motivo para Matricular-se			
		Qualidade da Escola	Custo da Escola ou Conveniência	Outros	Totais
Tipo de Matrícula	Tempo Integral	421	393	76	890
	Tempo Parcial	400	593	46	1.039
	Totais	821	986	122	1.929

    a. Desenvolva uma tabela de probabilidade associada para esses dados.
    b. Use as probabilidades marginais correspondentes à qualidade da escola, custo ou conveniência e outros para comentar a razão mais importante para alguém escolher a escola.
    c. Se o estudante optou por tempo integral, qual é a probabilidade de a qualidade ser a primeira razão para a escolha da escola?
    d. Se o estudante decidiu por tempo parcial, qual é a probabilidade de a qualidade ser a primeira razão para a escolha da escola?
    e. Admitamos que A denote o evento de um estudante estar em um curso de tempo integral e que B denote o evento de o estudante relacionar a qualidade da escola como a primeira razão para matricular-se. Os eventos A e B são independentes? Justifique sua resposta.

34. A tabela a seguir apresenta a distribuição dos tipos de sangue da população em geral (Hoxworth Blood Center, Cincinnati, Ohio, março de 2003).

	A	B	AB	O
Rh+	0,34	0,09	0,04	0,38
Rh–	0,06	0,02	0,01	0,06

a. Qual é a probabilidade de uma pessoa ter sangue do tipo O?
b. Qual a probabilidade de uma pessoa ser Rh–?
c. Qual é a probabilidade de uma pessoa ser Rh– sendo do grupo sangüíneo do tipo O?
d. Qual é a probabilidade de uma pessoa ter o tipo sangüíneo B sendo Rh+?
e. Qual é a probabilidade de, em um casal, ambos os cônjuges serem Rh–?
f. Qual é a probabilidade de, em um casal, ambos os cônjuges terem o tipo sangüíneo AB?

35. O U.S. Bureau of Labor Statistics colheu dados sobre a ocupação de trabalhadores cujas idades variavam de 25 a 64 anos. A tabela a seguir apresenta o número de trabalhadores e trabalhadoras (em milhões) em cada categoria de ocupação (*Statistical Abstract of the United States 2002*).

Ocupação	Homens	Mulheres
Área gerencial/profissional liberal	19.079	19.021
Área técnica/vendas/administrativa	11.079	19.315
Serviço	4.977	7.947
Produção de precisão	11.682	1.138
Operadores/manufatura/mão-de-obra	10.576	3.482
Agricultura/administração florestal/pesca	1.838	514

a. Desenvolva uma tabela de probabilidade associada.
b. Qual é a probabilidade de uma mulher trabalhadora ser gerente ou profissional liberal?
c. Qual é a probabilidade de um homem trabalhador ser da área de produção de precisão?
d. A ocupação independe de sexo? Justifique sua resposta com um cálculo de probabilidade.

36. Reggie Miller, do Indiana Pacers, é o melhor arremessador de lances livres da National Basketball Association, acertando 89% de seus arremessos (*USA Today*, 22 de janeiro de 2004). Suponha que no fim de um jogo de basquete, Reggie Miller sofra uma falta e se encarregue da cobrança de dois lances livres.

a. Qual é a probabilidade que ele tem de acertar ambos os arremessos?
b. Qual é a probabilidade que ele tem de acertar pelo menos um dos arremessos?
c. Qual é a probabilidade que ele tem de errar os dois arremessos?
d. No fim de um jogo de basquete, freqüentemente uma equipe comete falta em um jogador adversário a fim de parar o cronômetro do jogo. A estratégia habitual é cometer falta intencionalmente no pior arremessador de lances livres da outra equipe. Suponha que o pivô do Indiana Pacers acerte 58% de seus arremessos de lances livres. Calcule as probabilidades do pivô conforme as indicações nos itens (a), (b) e (c) e demonstre que cometer faltas intencionalmente no pivô do Indiana Pacers é uma estratégia melhor que cometer faltas intencionalmente em Reggie Miller.

37. Um agente de compras fez encomendas urgentes de determinada matéria-prima a dois diferentes fornecedores, *A* e *B*. Se nenhuma encomenda chegar em quatro dias, o processo de produção precisará ser interrompido até que pelo menos uma das encomendas chegue. A probabilidade de o fornecedor *A* poder entregar a matéria-prima em quatro dias é 0,55. A probabilidade de o fornecedor *B* poder entregar a matéria-prima em quatro dias é 0,35.

a. Qual é a probabilidade de ambos os fornecedores entregarem a matéria-prima em quatro dias? Uma vez que dois fornecedores distintos estão envolvidos, estamos inclinados a supor independência.
b. Qual é a probabilidade de pelo menos um fornecedor entregar a matéria-prima em quatro dias?
c. Qual é a probabilidade de o processo de produção estar paralisado em quatro dias em virtude da escassez da matéria-prima (ou seja, ambas as encomendas estarem atrasadas)?

38. A Minneapolis Heart Institute Foundation promoveu um estudo para determinar o benefício de fornecer tratamento de acompanhamento a pacientes que tiveram alta hospitalar depois do tratamento de um ataque cardíaco (*The Wall Street Journal*, 11 de novembro de 2002). Dos 2.060 pacientes, 1.070 não retornaram para o tratamento de acompanhamento e 990 o fizeram. Dentro de 24 meses, 14 dos pacientes que recebiam tratamento de acompanhamento morreram e 29 dos pacientes que não

recebiam tratamento de acompanhamento também morreram. Dentro de 54 meses, 20 dos pacientes que recebiam tratamento de acompanhamento morreram, e 49 dos pacientes que não recebiam tratamento de acompanhamento morreram.

a. Qual é a probabilidade de um paciente morrer dentro de 24 meses após a alta hospitalar para tratamento de um ataque cardíaco?
b. Usando os dados correspondentes aos 24 meses após a alta hospitalar, calcule as probabilidades condicionais de os pacientes que recebem e que não recebem tratamento de acompanhamento virem a morrer.
c. A probabilidade de morrer dentro de 24 meses após a alta hospitalar independe de a pessoa receber tratamento de acompanhamento? Explique.
d. Usando os dados correspondentes aos 54 meses, calcule as probabilidades condicionais de os pacientes que recebem e que não recebem tratamento de acompanhamento virem a morrer.
e. Você recomendaria a um amigo inscrever-se em um programa de tratamento de acompanhamento?

4.5 TEOREMA DE BAYES

Na discussão da probabilidade condicional, indicamos que revisar as probabilidades quando se obtêm novas informações é uma etapa importante da análise de probabilidades. Freqüentemente, iniciamos a análise com **estimativas da probabilidade inicial** ou **a priori** para eventos de interesse específico. Então, a partir de fontes como uma amostra, relatório especial ou teste de produto, obtemos informações adicionais sobre os eventos. Dadas essas novas informações, atualizamos os valores da probabilidade prévia calculando as probabilidades revisadas, chamadas **probabilidades a posteriori**. O **teorema de Bayes** constitui um meio de efetuarmos esses cálculos de probabilidade. As etapas desse processo de revisão de probabilidade são mostradas na Figura 4.9.

Como uma aplicação do teorema de Bayes, considere uma firma de manufatura que recebe remessas de peças e dois diferentes fornecedores. Digamos que A_1 denote o evento de uma peça ser proveniente do fornecedor 1 e A_2 denote o evento de a peça vir do fornecedor 2. Atualmente, 65% das peças compradas pela empresa são do fornecedor 1 e os restantes 35% são do fornecedor 2. Portanto, se uma peça for escolhida aleatoriamente, atribuiríamos as probabilidades iniciais $P(A_1) = 0{,}65$ e $P(A_2) = 0{,}35$.

Figura 4.9 Revisão da probabilidade usando o Teorema de Bayes

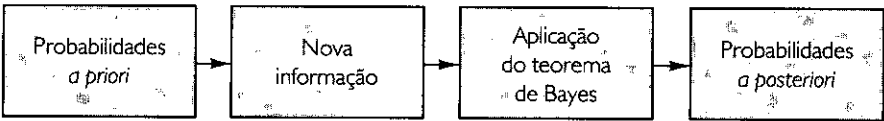



Tabela 4.6 Níveis históricos da qualidade de dois fornecedores

	Porcentagem de Peças Boas	Porcentagem de Peças Ruins
Fornecedor 1	98	2
Fornecedor 2	95	5

A qualidade das peças compradas varia de acordo com a fonte de fornecimento. Os dados históricos sugerem que as avaliações da qualidade dos dois fornecedores são similares às que são apresentadas na Tabela 4.6. Se admitirmos que B denota o evento de uma peça boa e R denota o evento de uma peça ruim, a informação contida na Tabela 4.6 nos oferece os seguintes valores de probabilidade condicional.

$$P(B \mid A_1) = 0{,}98 \; P(R \mid A_1) = 0{,}02$$
$$P(B \mid A_2) = 0{,}95 \; P(R \mid A_2) = 0{,}05$$

O diagrama em árvore da Figura 4.10 descreve o processo de a empresa receber uma peça de um dos dois fornecedores e depois descobrir que a peça é boa ou ruim como um experimento de duas etapas.

Notamos que são possíveis quatro resultados experimentais: dois correspondem ao fato de a peça ser boa e dois correspondem ao fato de a peça ser ruim.

Cada um dos resultados experimentais é a interseção de dois eventos, de forma que podemos usar a regra de multiplicação para calcular as probabilidades. Por exemplo,

$$P(A_1, B) = P(A_1 \cap B) = P(A_1)P(B \mid A_1)$$

O processo de cálculo dessas probabilidades associadas pode ser retratado por meio daquilo que chamamos árvore de probabilidades (veja a Figura 4.11). Da esquerda para a direita ao longo da árvore, as probabilidades correspondentes a cada uma das ramificações indicadas na etapa 1 são probabilidades *a priori*, e as probabilidades correspondentes a cada uma das ramificações indicadas na etapa 2 são probabilidades condicionais.

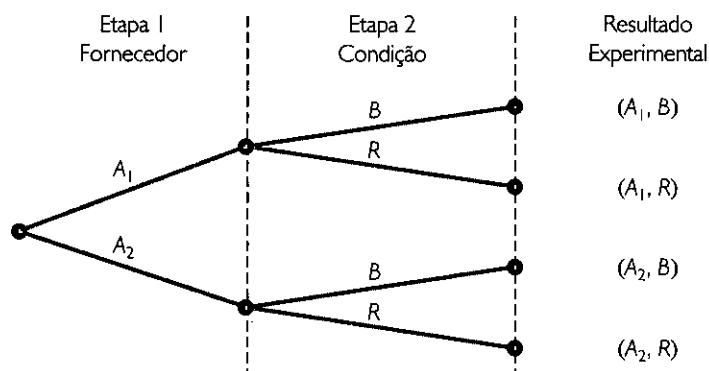
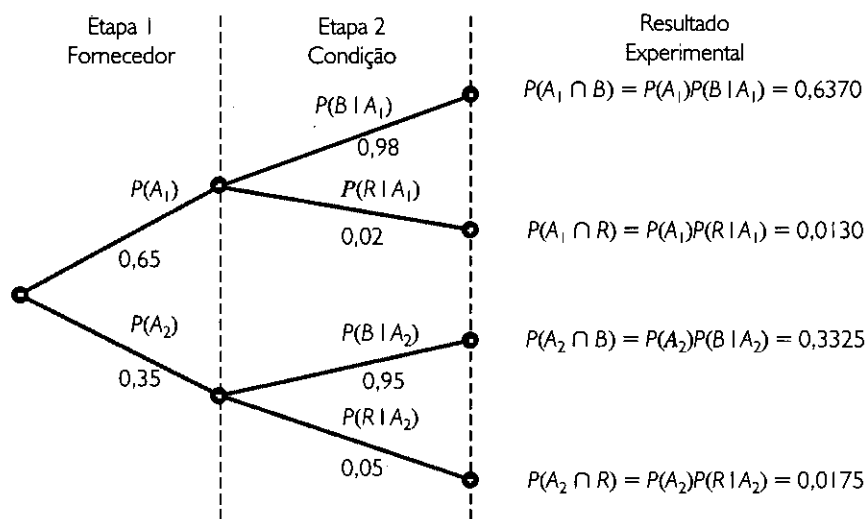
Figura 4.10 Diagrama em árvore correspondente ao exemplo dos dois fornecedores

Nota: A etapa 1 mostra que a peça vem de um dos dois fornecedores, e a etapa 2 mostra se a peça é boa ou ruim.

Figura 4.11 Árvore de probabilidades correspondente ao exemplo dos dois fornecedores

Para encontrar as probabilidades de cada resultado experimental, simplesmente multiplicamos as probabilidades nas ramificações que levam ao resultado. Cada uma dessas probabilidades associadas é mostrada na Figura 4.11, juntamente com as probabilidades conhecidas correspondentes a cada ramificação.

Suponha agora que as peças recebidas dos dois fornecedores são usadas no processo manufatureiro da firma e que uma máquina se quebre ao tentar processar uma peça ruim. Dada a informação de que a peça é ruim, qual é a probabilidade de ela ter vindo do fornecedor 1 e qual é a probabilidade de ela ter vindo do fornecedor 2? Com as informações contidas na árvore de probabilidades (Figura 4.11), o teorema de Bayes pode ser usado para responder a essas questões.

Admitindo que R denote o evento de a peça ser ruim, estamos à procura das probabilidades $P(A_1 \mid B)$ e $P(A_2 \mid B)$. Da lei da probabilidade condicional, sabemos que

$$P(A_2 \mid R) = \frac{P(A_1 \cap R)}{P(R)} \qquad \textbf{(4.14)}$$

Consultando a árvore de probabilidades, vemos que

$$P(A_1 \cap R) = P(A_1)P(R \mid A_1) \qquad \textbf{(4.15)}$$

Para encontrar $P(R)$, notamos que o evento R pode ocorrer somente de duas maneiras: $(A_1 \cap R)$ e $(A_2 \cap R)$. Portanto, temos

$$\begin{aligned} P(R) &= P(A_1 \cap R) + P(A_1 \cap R) \\ &= P(A_1)P(R \mid A_1) + P(A_2)P(R \mid A_2) \qquad \textbf{(4.16)} \end{aligned}$$

Substituindo os dados das Equações 4.15 e 4.16 na Equação 4.14 e escrevendo um resultado similar para $P(A_2 \mid R)$, obtemos o teorema de Bayes para o caso de dois eventos.

TEOREMA DE BAYES (CASO DE DOIS EVENTOS)

$$P(A_1 \mid R) = \frac{P(A_1)P(R \mid A_1)}{P(A_1)P(R \mid A_1) + P(A_2)P(R \mid A_2)} \qquad \textbf{(4.17)}$$

$$P(A_2 \mid R) = \frac{P(A_2)P(R \mid A_2)}{P(A_1)P(R \mid A_1) + P(A_2)P(R \mid A_2)} \qquad \textbf{(4.18)}$$

Credita-se ao reverendo Thomas Bayes (1702-1761), um ministro presbiteriano, o trabalho original que levou à versão do teorema de Bayes que usamos atualmente.

Usando a Equação 4.17 e os valores de probabilidade do exemplo, temos

$$\begin{aligned} P(A_1 \mid R) &= \frac{P(A_1)P(R \mid A_1)}{P(A_1)P(R \mid A_1) + P(A_2)P(R \mid A_2)} \\ &= \frac{(0{,}65)(0{,}02)}{(0{,}65)(0{,}02) + (0{,}35)(0{,}05)} = \frac{0{,}013}{0{,}013 + 0{,}0175} \\ &= \frac{0{,}013}{0{,}0305} = 0{,}4262 \end{aligned}$$

Além disso, usando a Equação 4.18, encontramos $P(A_2 \mid R)$.

$$\begin{aligned} P(A_2 \mid R) &= \frac{(0{,}35)(0{,}05)}{(0{,}65)(0{,}02) + (0{,}35)(0{,}05)} \\ &= \frac{0{,}175}{0{,}0130 + 0{,}175} = \frac{0{,}175}{0{,}305} = 0{,}5738 \end{aligned}$$

Note que nessa aplicação iniciamos com a probabilidade de 0,65 de que uma peça escolhida aleatoriamente tenha sido do fornecedor 1. Entretanto, dada a informação de que a peça é ruim, a probabilidade de que ela seja do fornecedor 1 cai para 0,4262. De fato, se a peça for ruim, ela tem uma chance maior que 50:50 de ter vindo do fornecedor 2: ou seja, $P(A_2 \mid R) = 0{,}5738$.

O teorema de Bayes é aplicável quando os eventos para os quais queremos calcular probabilidades *a posteriori* são mutuamente exclusivos e a união deles é o espaço amostral inteiro.[3] Para o caso de n eventos $A_1, A_2, \ldots, A_n$ mutuamente exclusivos, cuja união é o espaço amostral inteiro, o teorema de Bayes pode ser usado para calcular qualquer probabilidade *a posteriori* $P(A_1 \mid R)$, como mostramos aqui:

TEOREMA DE BAYES

$$P(A_i \mid R) = \frac{P(A_i)P(R \mid A_i)}{P(A_1)P(R \mid A_1) + P(A_2)P(R \mid A_2) + \cdots P(A_n)P(R \mid A_n)} \quad (4.19)$$

Com as probabilidades *a priori* $P(A_1), P(A_2), \ldots, P(A_n)$ e as probabilidades condicionais apropriadas $P(R \mid A_1), P(R \mid A_2), \ldots, P(R \mid A_n)$, pode-se usar a Equação 4.19 para calcular a probabilidade *a posteriori* dos eventos $A_1, A_2, \ldots, A_n$.

A Abordagem Tabular

Uma abordagem tabular é útil para se efetuarem os cálculos do teorema de Bayes. Esse tipo de abordagem é mostrado na Tabela 4.7, correspondente ao problema dos fornecedores de peças. Os cálculos lá mostrados são feitos nas seguintes etapas:

Etapa 1. Prepare as três colunas seguintes:
Coluna 1 – Os eventos A_i mutuamente exclusivos para os quais se desejam as probabilidades *a posteriori*.
Coluna 2 – As probabilidades *a priori* $P(A_i)$ dos eventos.
Coluna 3 – As probabilidades condicionais $P(R \mid A_i)$ da nova informação R dada para cada evento.

Etapa 2. Na coluna 4, calcule as probabilidades associadas $P(A_i \cap R)$ correspondentes a cada evento, e a nova informação R usando-se a lei da multiplicação. Essas probabilidades associadas são encontradas multiplicando-se as probabilidades iniciais da coluna 2 pelas probabilidades condicionais correspondentes na coluna 3; ou seja, $P(A_i \cap R) = P(A_i)P(R \mid A_i)$.

Etapa 3. Some as probabilidades associadas da coluna 4. A soma é a probabilidade da nova informação, $P(R)$. Desse modo, vemos na Tabela 4.7 que há uma probabilidade de 0,0130 de a peça ruim ter vindo do fornecedor 1 e uma probabilidade de 0,0175 de a peça ruim ter vindo do fornecedor 2. Desde que estas sejam as duas únicas maneiras pelas quais uma peça ruim pode ser obtida, a soma 0,0130 + 0,0175 mostra uma probabilidade global de 0,0305 de se encontrar uma peça ruim nas remessas conjuntas dos dois fornecedores.

Etapa 4. Na coluna 5, calcule as probabilidades *a posteriori* usando a relação básica de probabilidade condicional.

$$P(A_i \mid R) = \frac{P(A_i \cap R)}{P(B)}$$

Observe que as probabilidades associadas $P(A_i \cap R)$ estão na coluna 4 e que a probabilidade $P(R)$ é a soma da coluna 4.

Tabela 4.7 Abordagem tabular para cálculos do Teorema de Bayes referentes ao problema dos dois fornecedores

(1) Eventos	(2) Probabilidades *a Priori* $P(A_i)$	(3) Probabilidades Condicionais $P(R \mid A_i)$	(4) Probabilidades Associadas $P(A_i \cap R)$	(5) Probabilidades *a Posteriori* $P(A_i \mid R)$
A_i	0,65	0,02	0,0130	0,0130/0,0305 = 0,4262
A_l	0,35	0,05	0,0175	0,0175/0,0305 = 0,5738
A_2	1,00		$P(R)$ = 0,0305	1,0000

[3] Se a união dos eventos for o espaço amostral inteiro, diz-se que os eventos são *coletivamente exaustivos*.

NOTAS E COMENTÁRIOS

1. O teorema de Bayes é amplamente usado na análise de decisões. As probabilidades iniciais freqüentemente são estimativas subjetivas apresentadas por um tomador de decisões. As informações da amostra são obtidas e as probabilidades *a posteriori* são calculadas a fim de serem utilizadas na escolha da melhor decisão.
2. Um evento e seu complemento são mutuamente exclusivos, e sua união constitui o espaço amostral inteiro. Desse modo, o teorema de Bayes é sempre aplicável quando se quer calcular as probabilidades *a posteriori* de um evento e seu complemento.

Exercícios

Métodos

AUTOTESTE

39. As probabilidades *a priori* dos eventos A_1 e A_2 são $P(A_1) = 0{,}40$ e $P(A_2) = 0{,}60$. Sabe-se também que $P(A_1 \cap A_2) = 0$. Suponha que $P(R \mid A_1) = 0{,}20$ e $P(R \mid A_2) = 0{,}05$.
   a. A_1 e A_2 são mutuamente exclusivos? Explique.
   b. Calcule $P(A_1 \cap R)$ e $P(A_2 \cap R)$.
   c. Calcule $P(R)$.
   d. Aplique o teorema de Bayes para calcular $P(A_1 \mid R)$ e $P(A_2 \mid R)$.
40. As probabilidades iniciais dos eventos A_1, A_2 e A_3 são $P(A_1) = 0{,}20$, $P(A_2) = 0{,}50$ e $P(A_3) = 0{,}30$. As probabilidades condicionais do evento B, dados A_1, A_2 e A_3 são $P(R \mid A_1) = 0{,}50$, $P(R \mid A_2) = 0{,}40$ e $P(R \mid A_3) = 0{,}30$.
   a. Calcule $P(R \cap A_1)$, $P(R \cap A_2)$ e $P(R \cap A_3)$.
   b. Aplique o teorema de Bayes, a Equação 4.19, para calcular a probabilidade *a posteriori* $P(A_2 \mid R)$.
   c. Use a abordagem tabular para aplicar o teorema de Bayes ao cálculo de $P(A_1 \mid R)$, $P(A_2 \mid R)$ e $P(A_3 \mid R)$.

Aplicações

41. Uma firma de consultoria apresentou uma proposta para a execução de um grande projeto de pesquisa. A gerência da firma achava inicialmente que tinham uma chance de 50:50 de obter o projeto. No entanto, o órgão para o qual a proposta foi submetida solicitou subseqüentemente informações adicionais sobre a proposta apresentada. A experiência indica que para 75% das propostas bem-sucedidas e para 40% das propostas malsucedidas o órgão solicitara informações adicionais.
   a. Qual é a probabilidade *a priori* de a proposta ser bem-sucedida (isto é, antes do pedido de informações adicionais)?
   b. Qual é a probabilidade condicional de um pedido de informações adicionais, dado que a proposta seja, por fim, bem-sucedida?
   c. Calcule a probabilidade *a posteriori* de que a proposta seja bem-sucedida, dado um pedido de informações adicionais.

AUTOTESTE

42. Um banco local fez uma revisão de sua política de cartões de crédito com a intenção de cancelar alguns contratos de cartões. No passado, aproximadamente 5% dos detentores de cartões de crédito se tornaram inadimplentes, deixando o banco incapaz de cobrar o saldo devedor. Portanto, a gerência estabeleceu uma probabilidade *a priori* de 0,05 de que qualquer portador de cartão de crédito em particular se tornará inadimplente. O banco também descobriu que a probabilidade de os clientes que não são inadimplentes deixarem de efetuar um pagamento mensal é 0,20. Naturalmente, a probabilidade de os inadimplentes deixarem de efetuar um pagamento mensal é 1.
   a. Dado que o cliente tenha deixado de efetuar um ou mais pagamentos mensais, calcule a probabilidade *a posteriori* de que o cliente se torne inadimplente.
   b. O banco gostaria de cancelar o cartão de crédito se a probabilidade de um cliente tornar-se inadimplente for maior que 0,20. O banco deveria cancelar o cartão se o cliente deixar de efetuar um pagamento mensal? Por quê?
43. Carros pequenos têm um melhor desempenho quanto ao consumo de combustível por quilômetro, mas não são tão seguros quanto os carros maiores. Os carros pequenos são responsáveis por 18% dos veículos nas estradas, mas os acidentes envolvendo carros pequenos acarretaram 11.898 mortes durante

um ano recente (*Reader's Digest*, maio de 2000). Suponha que a probabilidade de um carro pequeno envolver-se em um acidente seja 0,18. A probabilidade de um acidente envolvendo um carro pequeno e que provoca uma morte é 0,128, e a probabilidade de um acidente não envolvendo um carro pequeno e que acarreta uma morte é 0,05. Suponha que você soube de um acidente envolvendo uma morte. Qual é a probabilidade de um carro pequeno estar envolvido nesse acidente? Suponha que a probabilidade de envolver-se em um acidente independa do tamanho do carro.

44. O American Council of Education divulgou que 47% dos calouros universitários colam graus e fazem pós-graduação em cinco anos (*Associated Press*, 6 de maio de 2002). Suponha que os registros do curso de pós-graduação mostrem que as mulheres compõem 50% dos estudantes que se graduaram em cinco anos, mas somente 45% dos estudantes que não se graduaram em cinco anos. Os estudantes que não se graduaram em cinco anos ou saíram da escola ou ainda não tinham concluído o curso.
    a. Se admitirmos que A_1 = o estudante que se graduou em cinco anos
       A_2 = o estudante que não se graduou em cinco anos
       M = o estudante é uma mulher
    Usando a informação dada, quais são os valores para $P(A_1)$, $P(A_2)$, $P(M \mid A_1)$ e $P(M \mid A_2)$?
    b. Qual é a probabilidade de uma mulher graduar-se dentro de cinco anos?
    c. Qual é a probabilidade de um homem graduar-se dentro de cinco anos?
    d. Dados os resultados anteriores, qual é a porcentagem de mulheres e qual é a porcentagem de homens calouros na universidade?

45. Em um artigo sobre o aumento dos investimentos, a revista *Money* relatou que os títulos de empresas de produtos farmacêuticos exibem fortes tendências a longo prazo e oferecem aos investidores um potencial incomparável para a obtenção de retornos volumosos e constantes. A Health Care Financing Administration do governo federal sustenta essa conclusão por meio de sua previsão de que os gastos anuais com medicamentos vendidos sob prescrição médica passarão dos US$ 117 bilhões em 2000 para US$ 366 bilhões em 2010. Muitas pessoas que têm mais de 65 anos recorrem fortemente aos medicamentos vendidos sob prescrição médica. Em relação a esse grupo, 82% tomam regularmente medicamentos vendidos com receita, 55% tomam regularmente três ou mais medicamentos vendidos com receita, e 40% usam atualmente cinco ou mais remédios vendidos com receita. Comparativamente, 49% das pessoas com menos de 65 anos tomam regularmente remédios vendidos com prescrição médica, e 17% tomam regularmente três ou mais remédios vendidos com receita e 28% usam cinco ou mais remédios vendidos com receita (*Money*, setembro de 2001). O Departamento do Censo norte-americano relata que da população de 281.421.906 nos Estados Unidos, 34.991.753 têm mais de 65 anos (U.S. Census Bureau, *Census 2000*).
    a. Calcule a probabilidade de uma pessoa nos Estados Unidos ter 65 anos ou mais.
    b. Calcule a probabilidade de uma pessoa tomar remédios com prescrição médica regularmente.
    c. Calcule a probabilidade de uma pessoa com 65 anos ou mais tomar cinco ou mais medicamentos vendidos com receita médica.
    d. Dado que uma pessoa use cinco ou mais medicamentos vendidos com receita médica, calcule a probabilidade de essa pessoa ter 65 anos ou mais.

Resumo

Neste capítulo, introduzimos os conceitos básicos de probabilidade e ilustramos como a análise das probabilidades pode ser usada para fornecer informações úteis para a tomada de decisões. Descrevemos como a probabilidade pode ser interpretada como a medida numérica da possibilidade de um evento ocorrer. Além disso, vimos que a probabilidade de um evento pode ser calculada somando-se as probabilidades dos resultados experimentais (pontos amostrais) que compreendem o evento ou usando-se as relações estabelecidas pelas leis de probabilidade da adição, da probabilidade condicional e da multiplicação. Para os casos em que informações adicionais estão disponíveis, mostramos como o teorema de Bayes pode ser usado para se obter probabilidades revisadas ou posteriores.

Glossário

Probabilidade Medida numérica da possibilidade de um evento ocorrer.

Experimento Processo que gera resultados bem definidos.

Espaço amostral Conjunto de todos os resultados experimentais.

Ponto amostral Elemento do espaço amostral. Um ponto amostral representa um resultado experimental.

Diagrama em árvore Representação gráfica que ajuda a visualizar um experimento de múltiplas etapas.

Requisitos básicos para a atribuição de probabilidades Dois requisitos que restringem a maneira pela qual se podem fazer atribuições de probabilidades: (a) Para cada resultado experimental E_i, devemos ter $0 \leq P(E_i) \leq 1$; (b) Considerando-se todos os resultados experimentais, devemos ter $P(E_1) + P(E_2) + \ldots + P(E_n) = 1{,}0$.

Método clássico Método de atribuir probabilidades que é apropriado quando todos os resultados experimentais são igualmente prováveis.

Método da freqüência relativa Método de atribuição de probabilidades que é apropriado quando há dados disponíveis para estimar a proporção do tempo que o resultado experimental ocorrerá se o experimento for repetido um grande número de vezes.

Método subjetivo Método de atribuição de probabilidades que se baseia no julgamento.

Evento Conjunto de pontos amostrais.

Complemento de *A* Evento que consiste em todos os pontos amostrais que não estão em A.

Diagrama de Venn Representação gráfica para exibir simbolicamente o espaço amostral e as operações que envolvem eventos na qual o espaço amostral é representado por um retângulo e os eventos são representados por círculos.

União de *A* e *B* Evento que contém todos os pontos amostrais que pertencem a A, B, ou a ambos. A união é denotada por $A \cup B$.

Interseção de *A* e *B* Evento que contém todos os pontos amostrais que pertencem tanto a A como a B. A interseção é denotada por $A \cap B$.

Lei da adição Lei de probabilidade usada para calcular a probabilidade da união de dois eventos. Ela é $P(A \cup B) = P(A) + P(B) - P(A \cap B)$. Para eventos mutuamente exclusivos, $P(A \cap B) = 0$; nesse caso, a lei da adição se reduz a $P(A \cup B) = P(A) + P(B)$.

Eventos mutuamente exclusivos Eventos que não têm pontos amostrais em comum, ou seja, $A \cap B$ é vazia e $P(A \cap B) = 0$.

Probabilidade condicional A probabilidade de um evento, dado que outro evento já tenha ocorrido. A probabilidade condicional de A, dado B, é $P(A \mid B) = P(A \cap B)/P(B)$.

Probabilidade associada A probabilidade de haver dois eventos e ambos ocorrerem; ou seja, a probabilidade da interseção de dois eventos.

Probabilidade marginal Os valores situados nas margens de uma tabela de probabilidade associada que fornecem as probabilidades de cada evento separadamente.

Eventos independentes Dois eventos A e B, em que $P(A \mid B) = P(A)$ ou $P(B \mid A) = P(B)$; ou seja, os eventos não têm nenhuma influência mútua.

Lei da multiplicação Lei de probabilidade usada para calcular a probabilidade da interseção de dois eventos. Ela é $P(A \cap B) = P(B)P(A \mid B)$ ou $P(A \cap B) = P(A)P(B \mid A)$. Para eventos independentes, ela se reduz a $P(A \cap B) = P(A)P(B)$.

Probabilidades *a priori* Estimativas iniciais das probabilidades dos eventos.

Probabilidades *a posteriori* Probabilidades revisadas dos eventos, baseadas em informações adicionais.

Teorema de Bayes Método usado para calcular probabilidades *a posteriori*.

Fórmulas-Chave

Regra de Contagem de Combinações

$$C_n^N = \binom{N}{n} = \frac{N!}{n!(N-n)!} \tag{4.1}$$

Regra de Contagem de Permutações

$$P_n^N = n!\binom{N}{n} = \frac{N!}{(N-n)!} \tag{4.2}$$

Cálculo da Probabilidade Usando o Complemento

$$P(A) = 1 - P(A^c) \tag{4.5}$$

Lei da Adição

$$P(A \cup B) = P(A) + P(B) - P(A \cap B) \tag{4.6}$$

Probabilidade Condicional

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)} \tag{4.7}$$

$$P(B \mid A) = \frac{P(A \cap B)}{P(A)} \tag{4.8}$$

Lei da Multiplicação

$$P(A \cap B) = P(B)P(A \mid B) \tag{4.11}$$

$$P(A \cap B) = P(A)P(B \mid A) \tag{4.12}$$

Lei da Multiplicação de Eventos Independentes

$$P(A \cap B) = P(A)P(B) \tag{4.13}$$

Teorema de Bayes

$$P(A_i \mid B) = \frac{P(A_i)P(B \mid A_i)}{P(A_1)P(B \mid A_1) + P(A_2)P(B \mid A_2) + \cdots + P(A_n)P(B \mid A_n)} \tag{4.19}$$

Exercícios Suplementares

46. Em uma pesquisa de opinião realizada pela *Business Week*/Harris Poll, 1.035 pessoas adultas foram solicitadas a dizer qual postura elas tinham em relação aos negócios (*Business Week*, 11 de setembro de 2000). Foi feita a seguinte pergunta: "Como você avaliaria as grandes empresas norte-americanas em termos de produzirem bons produtos e competirem em um ambiente globalizado?" As respostas foram: excelente – 18%, ótimo – 50%, razoável – 26%, ruim – 5% e não sei/nenhuma resposta – 1%.
    a. Qual é a probabilidade de uma pessoa consultada ter avaliado as empresas norte-americanas como ótimas ou excelentes?
    b. Quantas pessoas consultadas avaliaram as empresas norte-americanas como ruins?
    c. Quantas pessoas consultadas não souberam dar uma resposta ou não responderam?

47. Um gerente financeiro fez dois novos investimentos: um na indústria petrolífera e um em títulos municipais. Após o período de um ano, cada um dos investimentos será classificado como bem-sucedido ou malsucedido. Considere a realização dos dois investimentos como um experimento.
    a. Quantos pontos amostrais existem para esse experimento?
    b. Apresente um diagrama em árvore e relacione os pontos amostrais.
    c. Admitamos que O = o evento de o investimento na indústria petrolífera ser bem-sucedido e M = o evento de o investimento em títulos municipais ser bem-sucedido. Relacione os pontos amostrais em O e em M.
    d. Relacione os pontos amostrais na união dos eventos $(O \cup M)$.
    e. Relacione os pontos amostrais na interseção dos eventos $(O \cap M)$.
    f. Os eventos O e M são mutuamente exclusivos? Explique.

48. No início de 2003 o presidente Bush propôs eliminar a tributação dos dividendos pagos a acionistas afirmando que se tratava de dupla tributação. As corporações pagam impostos sobre os rendimentos que mais tarde são pagos em dividendos. Em uma pesquisa de opinião de 671 norte-americanos, a TechnoMetrica Market Intelligence descobriu que 47% dos entrevistados eram favoráveis à proposta, 44% se opunham e 9% não tinham certeza (*Investor's Business Daily*, 13 de janeiro de 2003). Ao examinar as respostas baseando-se na opção partidária, a pesquisa mostrou que 29% dos democratas eram favoráveis, 64% dos republicanos eram favoráveis e 48% dos independentes eram favoráveis.
    a. Quantos dos entrevistados eram favoráveis à eliminação da tributação dos dividendos?
    b. Qual é a probabilidade condicional das pessoas favoráveis à proposta, uma vez que a pessoa entrevistada é um democrata?
    c. A filiação partidária independe de alguém ser favorável à proposta?
    d. Se presumirmos que as respostas das pessoas foram coerentes com seus próprios interesses pessoais, qual grupo você acha que se beneficiaria mais com a aprovação da proposta?

49. Um estudo de 31 mil internações hospitalares no estado de Nova York descobriu que 4% das internações acarretavam lesões causadas pelo tratamento. Um sétimo dessas lesões resultou em morte e um quarto delas foi causado por negligência. Foram ajuizados pedidos de indenização por imperícia médica em um de cada 7,5 casos envolvendo negligência, e indenizações foram pagas em um de cada dois pedidos.
   a. Qual é a probabilidade de uma pessoa internada em um hospital sofrer uma lesão em razão da negligência?
   b. Qual é a probabilidade de uma pessoa internada em um hospital morrer em conseqüência de uma lesão causada pelo tratamento?
   c. No caso de uma lesão causada por negligência, qual é a probabilidade de uma reivindicação de indenização por imperícia médica ser paga?
50. Uma pesquisa feita por telefone para determinar a reação dos telespectadores a um novo programa de televisão obteve os seguintes dados.

Avaliação	Freqüência
Fraco	4
Abaixo da Média	8
Médio	11
Acima da Média	14
Excelente	13

   a. Qual é a probabilidade de um telespectador escolhido aleatoriamente avaliar que o novo programa é médio ou melhor?
   b. Qual é a probabilidade de um telespectador escolhido aleatoriamente avaliar que o novo programa está abaixo da média ou pior?
51. A seguinte tabulação cruzada apresenta a renda familiar por nível educacional do chefe da família (*Statistical Abstract of the United State 2002*).

	Renda Familiar (US$1.000s)					
Nível Educacional	**Abaixo de 25**	**25,0–49,9**	**50,0–74,9**	**75,0–99,9**	**100 ou mais**	**Total**
Sem Diploma do Ensino Médio	9.285	4.093	1.589	541	354	15.862
Com Diploma do Ensino Médio	10.150	9.821	6.050	2.737	2.028	30.786
Universitário Incompleto	6.011	8.221	5.813	3.215	3.120	26.380
Bacharelado	2.138	3.985	3.952	2.698	4.748	17.521
Superior ao Bacharelado	813	1.497	1.815	1.589	3.765	9.479
Total	28.397	27.617	19.219	10.780	14.015	100.028

   a. Desenvolva uma tabela de probabilidade associada.
   b. Qual é a probabilidade de um chefe de família não ter diploma de ensino médio?
   c. Qual é a probabilidade de um chefe de família ter grau de bacharel ou superior?
   d. Qual é a probabilidade de uma família dirigida por alguém que tem o grau de bacharel ganhar US$ 100 mil ou mais?
   e. Qual é a probabilidade de uma família ter renda abaixo de US$ 25 mil?
   f. Qual é a probabilidade de uma família dirigida por alguém que tem o grau de bacharel ganhar menos de US$ 25 mil?
   g. A renda familiar independe do nível educacional?
52. Uma pesquisa dos novos matriculandos no curso de MBA da GMAC forneceu os seguintes dados, correspondentes a 2.018 alunos.

		Inscritos em mais de uma Escola	
		Sim	**Não**
Faixa Etária	23 anos ou menos	207	201
	24–26	299	379
	27–30	185	268
	31–35	66	193
	36 anos ou mais	51	169

a. Considerando um estudante de MBA escolhido aleatoriamente, prepare uma tabela de probabilidade associada do experimento que consiste em observar a idade do aluno e se ele se inscreveu em uma ou mais escolas.
b. Qual é a probabilidade de um candidato escolhido aleatoriamente ter 23 anos ou menos?
c. Qual é a probabilidade de um candidato escolhido aleatoriamente ter mais de 26 anos?
d. Qual é a probabilidade de um candidato escolhido aleatoriamente ter-se inscrito em mais de uma escola?

53. Consulte novamente os dados da pesquisa dos novos matriculandos na GMAC apresentados no Exercício 52.
a. Dado que uma pessoa se inscreva em mais de uma escola, qual é a probabilidade de a pessoa ter entre 24 e 26 anos?
b. Dado que uma pessoa esteja na faixa etária de 36 anos ou mais, qual é a probabilidade de ela se inscrever em mais de uma escola?
c. Qual é a probabilidade de uma pessoa ter idade entre 24 e 26 anos ou ter-se inscrito em mais de uma escola?
d. Suponha que saibamos que uma pessoa se inscreveu somente em uma escola. Qual é a probabilidade de a pessoa ter 31 anos ou mais?
e. O número de escolas em que se os estudantes se inscrevem independe de idade? Explique.

54. Uma pesquisa de opinião da IBD/TIPP, realizada com o objetivo de saber qual era a postura das pessoas em relação aos investimentos e aposentadoria (*Investor's Business Daily*, 5 de maio de 2000), perguntou a homens e mulheres qual a importância do nível de risco existente na escolha de um investimento para aposentadoria. A tabela de probabilidade associada seguinte foi construída a partir dos dados produzidos. "Importante" significa que a pessoa consultada disse que o nível de risco era importante ou muito importante.

	Homens	Mulheres	Total
Importante	0,22	0,27	0,49
Não Importante	0,28	0,23	0,51
Total	0,50	0,50	1,00

a. Qual é a probabilidade de uma pessoa consultada na pesquisa dizer que o nível de risco é importante?
b. Qual é a probabilidade de um homem consultado na pesquisa dizer que o nível de risco é importante?
c. Qual é a probabilidade de uma mulher consultada na pesquisa dizer que o nível de risco é importante?
d. O nível de risco independe do sexo da pessoa consultada? Por quê?
e. As posturas de homens e mulheres diferem quanto ao risco?

55. Uma grande empresa de bens de consumo veiculou um anúncio de televisão de um de seus produtos de limpeza. Com base em pesquisa, foram atribuídas probabilidades aos seguintes eventos.

B = pessoas que compraram o produto
S = pessoas que se lembram de ter visto o anúncio
$B \cap S$ = pessoas que compraram o produto e que se lembram de ter visto o anúncio

As probabilidades atribuídas foram $P(B) = 0{,}20$, $P(S) = 0{,}40$ e $P(B \cap S) = 0{,}12$.
a. Qual é a probabilidade de uma pessoa comprar o produto por se lembrar de ter visto o anúncio? Ver o anúncio aumenta a probabilidade de a pessoa comprar o produto? No papel de tomador de decisões, você recomendaria prosseguir com a anúncio (supondo que o custo seja razoável)?
b. Suponha que as pessoas que não compram o produto de limpeza dessa empresa compram-no de seus concorrentes. Qual seria sua estimativa da fatia de mercado da empresa? Você acredita que continuar com o anúncio aumentaria a fatia de mercado da empresa? Por quê?
c. A empresa experimentou também outro anúncio e atribuiu a ele os valores $P(S) = 0{,}30$ e $P(B \cap S) = 0{,}10$. Qual é a $P(B \mid S)$ desse outro anúncio? Qual anúncio lhe parece ter maior efeito sobre as compras efetuadas pelos clientes?

56. A Cooper Realty é uma pequena empresa imobiliária localizada em Albany, Nova York, especializada principalmente em intermediar a venda de residências. Recentemente, eles se interessaram em determinar a probabilidade de uma das residências de sua relação de imóveis ser vendida dentro de certo número de dias. Uma análise de vendas da empresa de 800 casas nos anos anteriores forneceu os seguintes dados.

		Dias de Permanência na Lista Até Ser Vendida			
		Abaixo de 30	31–90	Acima de 90	Total
Preços de Oferta Inicial	Abaixo de $150.000	50	40	10	100
	$150.000–$199.999	20	150	80	250
	$200.000–$250.000	20	280	100	400
	Acima de $250.000	10	30	10	50
	Total	100	500	200	800

a. Se A for definido como o evento de uma casa permanecer na lista de imóveis mais de 90 dias antes de ser vendida, estime a probabilidade de A.
b. Se B for definido como o evento de o preço de oferta inicial ser abaixo de US$ 150 mil, estime a probabilidade de B.
c. Qual é a probabilidade de $A \cap B$?
d. Supondo que um contrato para intermediar a venda de uma casa acaba de ser assinado, com um preço de oferta inicial inferior a US$ 150 mil, qual é a probabilidade de a casa exigir mais de 90 dias para que a Cooper Realty efetue a venda?
e. Os eventos A e B são independentes?

57. Uma empresa estudou o número de *lost-time accidents* (LTA)[4] que ocorrem em sua planta industrial em Brownsville, Texas. Os registros históricos mostram que 6% dos empregados sofreram LTA no ano passado. A gerência acredita que um programa especial de segurança reduzirá esse tipo de acidente para 5% durante o ano atual. Além disso, ela estima que 15% dos empregados que sofreram *lost-time accidents* no ano passado voltarão a sofrê-los no ano atual.
    a. Qual porcentagem dos empregados sofrerá *lost-time accidents* em ambos os anos?
    b. Qual porcentagem dos empregados sofrerá pelo menos um *lost-time accident* durante o período de dois anos?
58. A equipe de auditoria do IRS – *Internal Revenue Service* (Departamento da Receita Federal) de Dallas, preocupada em identificar declarações do imposto de renda potencialmente fraudulentas, acredita que a probabilidade de descobrir uma declaração fraudulenta, na hipótese de a declaração conter deduções de contribuições que ultrapassem o padrão do IRS, é de 0,20. Desde que as deduções de contribuições não ultrapassem o padrão do IRS, a probabilidade de ocorrência de uma declaração fraudulenta cai para 0,02. Se 8% de todas as declarações ultrapassarem o padrão do IRS para deduções em razão das contribuições efetuadas, qual é a melhor estimativa da porcentagem de declarações fraudulentas?
59. Uma companhia petrolífera comprou os direitos de prospecção de petróleo em uma área territorial no Alasca. Estudos geológicos preliminares atribuíram as seguintes probabilidades iniciais:

$$P(\text{petróleo de alta qualidade}) = 0{,}50$$
$$P(\text{petróleo de média qualidade}) = 0{,}20$$
$$P(\text{nenhum petróleo}) = 0{,}30$$

    a. Qual é a probabilidade de encontrarem petróleo?
    b. Depois de perfurarem 60,96 metros no primeiro poço, foi realizado um exame do solo. As probabilidades de encontrarem um tipo de solo em particular identificado pelo exame são apresentadas a seguir:

$$P(\text{solo} \mid \text{petróleo de alta qualidade}) = 0{,}20$$
$$P(\text{solo} \mid \text{petróleo de média qualidade}) = 0{,}80$$
$$P(\text{solo} \mid \text{sem petróleo}) = 0{,}20$$

Como a empresa deve interpretar o exame do solo? Quais são as probabilidades revisadas e qual é a nova probabilidade de encontrarem petróleo?

60. Empresas que fazem negócios pela internet freqüentemente podem obter informações de probabilidade sobre visitantes do *website* a partir de sites visitados anteriormente. O artigo "Internet Marketing" (*Interfaces*, março/abril de 2001) descreveu como os dados de *clickstream* em sites visi-

[4] NT: Acidente ocupacional ou doença que impede a uma pessoa retornar ao trabalho no dia (ou turno) seguinte. Literalmente, "um acidente que faz perder tempo".

tados poderiam ser usados em conjunto com um esquema de atualização bayesiano para determinar o sexo do visitante de um site. A Par Fore criou um site para comercializar equipamentos e vestuário para a prática do golfe. A gerência queria que determinada oferta fosse apresentada a visitantes do sexo feminino e uma oferta diferente fosse apresentada a visitantes do sexo masculino. A partir de uma amostra de visitas anteriores ao *website*, a gerência soube que 60% dos visitantes da ParFore.com eram homens e 40%, mulheres.

a. Qual é a probabilidade *a priori* de o próximo visitante do site ser uma mulher?
b. Suponha que você saiba que o visitante atual da ParFore.com visitou anteriormente o site da Dillard e que é três vezes mais provável que mulheres visitem o site da Dillard do que homens. Qual é a probabilidade revisada de o visitante atual da ParFore.com ser uma mulher? Você deve exibir a oferta que atrai mais as visitantes do sexo feminino ou a que atrai mais os visitantes do sexo masculino?

Estudo de Caso – Os Juízes do Condado de Hamilton

Os juízes do Condado de Hamilton examinam milhares de processos por ano. Na imensa maioria das causas decididas, o veredicto se mantém. Entretanto, alguns casos interpõem apelação e, dos que interpõem apelação, alguns são revertidos. Kristen DelGuzzi, do jornal *The Cincinnati Enquirer*, realizou um estudo dos processos julgados pelos juízes do Condado de Hamilton ao longo de um período de três anos. A Tabela 4.8 apresenta os resultados de 182.908 processos julgados (resolvidos) pelos 38 juízes da Common Pleas Court, da Domestic Relations Court e da Municipal Court.

Dois dos juízes, Dinkelacker e Hogan, não serviram no mesmo tribunal durante o período de três anos inteiro.

A finalidade do estudo promovido pelo jornal foi avaliar o desempenho dos juízes. Freqüentemente, as apelações ocorrem em conseqüência de erros cometidos pelos juízes, e o jornal queria saber quais juízes realizavam um bom trabalho e quais cometiam demasiados erros. Você é convocado para auxiliar na análise dos dados. Use o seu conhecimento das probabilidades e das probabilidades condicionais para ajudar a avaliar os juízes. Talvez você também seja capaz de analisar a probabilidade de apelação e de reversão de veredictos nos processos encaminhados pelos diferentes tribunais.

Relatório Administrativo

Prepare um relatório com sua avaliação dos juízes. Inclua também uma análise da probabilidade da apelação e de reversão de veredictos nos três tribunais. No mínimo, seu relatório deve incluir o seguinte:

1. A probabilidade de os processos sofrerem apelação e veredictos serem revertidos nos três diferentes tribunais.
2. A probabilidade de um processo sofrer apelação em relação a cada juiz.
3. A probabilidade de um processo sofrer reversão do veredicto em relação a cada juiz.
4. A probabilidade de reversão, dada uma apelação, em relação a cada juiz.
5. Avalie os juízes dentro de cada tribunal. Declare os critérios que usou e apresente o fundamento lógico para sua escolha.

Tabela 4.8 Total de causas decididas, que sofreram apelação e que tiveram reversão do veredicto nos tribunais do Condado de Hamilton

ARQUIVO DA INTERNET
Judge

Common Pleas Court

Juiz	Total de Causas Decididas	Causas que Sofreram Apelação	Causas que Tiveram Reversão do Veredicto
Fred Cartolano	3.037	137	12
Thomas Crush	3.372	119	10
Patrick Dinkelacker	1.258	44	8
Timothy Hogan	1.954	60	7
Robert Kraft	3.138	127	7
William Mathews	2.264	91	18
William Morrissey	3.032	121	22
Norbert Nadel	2.959	131	20
Arthur Ney. Jr.	3.219	125	14
Richard Niehaus	3.353	137	16
Thomas Nurre	3.000	121	6
John O'Connor	2.969	129	12
Robert Ruehlman	3.205	145	18
J. Howard Sundermann	955	60	10
Ann Marie Tracey	3.141	127	13
Ralph Winkler	3.089	88	6
Total	43.945	1762	199

Domestic Relations Court

Juiz	Total de Causas Decididas	Causas que Sofreram Apelação	Causas que Tiveram Reversão do Veredicto
Penelope Cunningham	2.729	7	1
Patrick Dinkelacker	6.001	19	4
Deborah Gaines	8.799	48	9
Ronald Panioto	12.970	32	3
Total	30.499	106	17

Municipal Court

Juiz	Total de Causas Decididas	Causas que Sofreram Apelação	Causas que Tiveram Reversão do Veredicto
Mike Allen	6.149	43	4
Nadine Allen	7.812	34	6
Timothy Black	7.954	41	6
David Davis	7.736	43	5
Leslie Isaiah Gaines	5.282	35	13
Karla Grady	5.253	6	0
Deidra Hair	2.532	5	0
Dennis Helmick	7.900	29	5
Timothy Hogan	2.308	13	2
James Patrick Kenney	2.798	6	1
Joseph Luebbers	4.698	25	8
William Mallory	8.277	38	9
Melba Marsh	8.219	34	7
Beth Mattingly	2.971	13	1
Albert Mestemaker	4.975	28	9
Mark Painter	2.239	7	3
Jack Rosen	7.790	41	13
Mark Schweikert	5.403	33	6
David Stockdale	5.371	22	4
John A. West	2.797	4	2
Total	108.464	500	104

CAPÍTULO 5

Distribuições Discretas de Probabilidade

ESTATÍSTICA NA PRÁTICA

CITIBANK*
Long Island City, Nova York

O Citibank, principal subsidiária do Citigroup, Inc., fornece ampla gama de serviços financeiros (por exemplo, contas correntes e contas de poupança, empréstimos e hipotecas, serviços de seguros e de investimentos), por meio da estrutura estratégica exclusiva para prestar esses serviços, denominada Citibanking. Essa estrutura vincula uma identidade de marca sólida, ofertas de produtos consistentes e serviços de qualidade ao cliente por todo o mundo. O Citibanking permite ao cliente gerenciar seu dinheiro a qualquer hora, em qualquer lugar e de acordo com sua preferência. Quer necessite poupar para o futuro, quer necessite fazer empréstimos imediatos, você pode fazer tudo isso no Citibank.

Os caixas automáticos de última geração do Citibanking, localizados nos Centros Bancários Citicard (CBCs), possibilitam aos usuários realizar todos os serviços bancários 24 horas por dia, sete dias por semana. Mais de 150 diferentes funções bancárias, que variam de depósitos à gestão de investimentos, podem ser executadas com facilidade. Os caixas automáticos do Citibanking são muito mais que simples máquinas de dinheiro, a tal ponto que os usuários os usam para 80% de suas transações.

Cada caixa automático do Citibank opera como um sistema de fila de espera, e os clientes que buscam serviços chegam aleatoriamente. Se todos estiverem ocupados, os clientes que chegam esperam em fila.

* Os autores agradecem a Ms. Stacey Karter, do Citibank, por fornecer esta "Estatística na Prática".

Estudos periódicos de capacidade dos caixas são utilizados para analisar o tempo de espera dos clientes e determinar se caixas adicionais são necessárias.

Os dados coletados pelo Citibank mostraram que as chegadas aleatórias de clientes seguiam uma distribuição de probabilidade conhecida como distribuição de Poisson. Usando a distribuição de Poisson, o Citibank pode calcular probabilidades relativas ao número de clientes que chegam a um caixa durante qualquer período e tomar decisões quanto ao número de caixas automáticos necessários.

Por exemplo, seja x igual ao número de clientes que chegam durante o período de um minuto, e vamos supor que um caixa em particular tenha uma taxa média de chegada de dois clientes por minuto, a tabela seguinte mostra as probabilidades relativas ao número de clientes que chegam durante o período de um minuto.

x	Probabilidade
0	0,1353
1	0,2707
2	0,2707
3	0,1804
4	0,0902
5 ou mais	0,0527

As distribuições de probabilidade discretas, como as usadas pelo Citibank, são o assunto deste capítulo. Além da distribuição de Poisson, você aprenderá a respeito das distribuições binomiais e hipergeométricas e como elas podem ser usadas para fornecer informações úteis de probabilidade.

Neste capítulo, continuamos o estudo da probabilidade, introduzindo os conceitos de variáveis aleatórias e de distribuições de probabilidade. O foco deste capítulo são as distribuições de probabilidade discretas. Serão abordadas de maneira especial três distribuições de probabilidade discretas: a binomial, a de Poisson e a hipergeométrica.

5.1 VARIÁVEIS ALEATÓRIAS

Variáveis aleatórias devem assumir valores numéricos.

No Capítulo 4, definimos o conceito de experimento e seus resultados experimentais concomitantes. Uma variável aleatória fornece um meio para se descrever resultados experimentais usando-se valores numéricos.

VARIÁVEL ALEATÓRIA

Uma **variável aleatória** é uma descrição numérica do resultado de um experimento.

Com efeito, uma variável aleatória associa um valor numérico a cada resultado experimental possível. O valor numérico da variável aleatória em particular depende do resultado do experimento. Uma variável aleatória pode ser classificada como *discreta* ou *contínua*, dependendo dos valores numéricos que ela assume.

Variáveis Aleatórias Discretas

Uma variável aleatória que pode assumir tanto um número finito de valores como uma seqüência infinita de valores – tais como 0, 1, 2, ... – é denominada **variável aleatória discreta**. Por exemplo, considere o experimento de um contador que presta o exame público para perito-contador (*certified public accountant* – CPA). O exame é composto de quatro partes. Podemos definir uma variável aleatória como x = o número de partes em que ele foi aprovado no exame CPA. Trata-se de uma variável aleatória discreta porque ela pode assumir o número finito de valores 0, 1, 2, 3 ou 4.

Como outro exemplo de variável aleatória discreta, considere o experimento de carros que chegam a um posto de pedágio. A variável aleatória de interesse é x = o número de carros que chegam durante o período de um dia. Os valores possíveis de x vêm da seqüência de números inteiros 0, 1, 2 e assim por diante. Portanto, x é uma variável aleatória discreta que assume um dos valores dessa seqüência infinita. Embora muitos experimentos tenham resultados que são naturalmente descritos por valores numéricos, outros não o são. Por exemplo, uma das questões de uma pesquisa pode solicitar a um indivíduo que relembre a men-

sagem de um recente comercial de televisão. Esse experimento teria dois resultados possíveis: o indivíduo não é capaz de lembrar-se da mensagem e o indivíduo é capaz de recordar-se da mensagem. Podemos ainda descrever esses resultados experimentais numericamente definindo-se a variável aleatória discreta x da seguinte maneira: seja $x = 0$ se o indivíduo não consegue lembrar-se da mensagem, e $x = 1$ se o indivíduo consegue relembrar da mensagem. Os valores numéricos dessa variável aleatória são arbitrários (poderíamos usar 5 e 10), mas elas são aceitáveis em termos da definição de variável aleatória – a saber, x é uma variável aleatória porque fornece uma descrição numérica do resultado do experimento.

A Tabela 5.1 fornece exemplos adicionais de variáveis aleatórias discretas. Note que, em cada exemplo, a variável aleatória discreta assume um número finito de valores ou uma seqüência infinita de valores, tais como 0, 1, 2,... Variáveis aleatórias discretas desses tipos são discutidas em detalhe neste capítulo.

Tabela 5.1 Exemplos de variáveis aleatórias discretas

Experimento	Variável Aleatória (x)	Valores Possíveis para a Variável Aleatória
Contatar cinco clientes	Número de clientes que colocam um pedido de compra	0, 1 2, 3, 4, 5
Inspecionar um embarque de 50 rádios	Número de rádios defeituosos	0, 1, 2, ..., 49, 50
Operar um restaurante durante um dia	Número de clientes	0, 1, 2, 3, ...
Vender um automóvel	Gênero do cliente	0 se for masculino; 1 se for feminino

Variáveis Aleatórias Contínuas

Uma variável aleatória que pode assumir qualquer valor numérico em um intervalo ou em uma coleção de intervalos é chamada **variável aleatória contínua**. Resultados experimentais que se baseiam em escalas de medidas como tempo, peso, distância e temperatura podem ser descritos por meio de variáveis aleatórias contínuas. Por exemplo, considere o experimento de monitoração das chamadas telefônicas feitas ao escritório de reclamação de seguros de uma importante companhia de seguros. Suponha que a variável aleatória de interesse seja $x =$ o tempo em minutos entre as chamadas consecutivas. Essa variável aleatória pode assumir qualquer valor no intervalo $x \geq 0$. Realmente, um número infinito de valores é possível para x, incluindo valores como 1,26 minuto, 2,751 minutos, 4,3333 minutos e assim por diante. Como outro exemplo, considere um trecho de 144 km da estrada de rodagem interestadual I-75 ao norte de Atlanta, Geórgia. Para um serviço de emergência de ambulâncias localizado em Atlanta, podemos definir a variável aleatória como $x =$ o número de quilômetros até o local do próximo acidente de trânsito ao longo desse trecho da I-75. Nesse caso, x seria uma variável aleatória contínua que assume qualquer valor no intervalo $0 \leq x \leq 90$. Exemplos adicionais de variáveis aleatórias contínuas estão listados na Tabela 5.2. Note que cada exemplo descreve uma variável aleatória que pode assumir qualquer valor em um intervalo de valores. As variáveis aleatórias contínuas e suas distribuições de probabilidade serão o assunto do Capítulo 6.

Tabela 5.2 Exemplos de variáveis aleatórias contínuas

Experimento	Variável Aleatória (x)	Valores Possíveis para a Variável Aleatória
Operar um banco	Tempo em minutos entre as chegadas dos clientes	$x \geq 0$
Encher uma lata de refrigerante (máx. = 343 mL)	Quantidade em mL	$0 \leq x \leq 343$
Construir uma nova biblioteca	Porcentagem de conclusão do projeto depois de seis meses	$0 \leq x \leq 100$
Testar um novo processo químico	A temperatura quando ocorre a reação desejada (mín. 65 °C; máx. 100 °C)	$65° \leq x \leq 100°$

NOTAS E COMENTÁRIOS

Um modo de determinar se uma variável aleatória é discreta ou contínua é pensar nos valores da variável aleatória como pontos sobre um segmento de reta. Escolha dois pontos que representam os valores da variável aleatória. Se todo o segmento de reta entre os dois pontos também representa possíveis valores para a variável aleatória, então a variável aleatória é contínua.

Exercícios

Métodos

AUTOTESTE

1. Considere o experimento de jogar uma moeda duas vezes.
   a. Liste os resultados experimentais.
   b. Defina uma variável aleatória que represente o número de coroas que ocorrem nos dois arremessos.
   c. Mostre qual valor a variável aleatória assumiria para cada um dos resultados experimentais.
   d. A variável aleatória é discreta ou contínua?
2. Considere o experimento de um trabalhador que monta um produto.
   a. Defina uma variável aleatória que represente o tempo necessário em minutos para montar o produto.
   b. Quais valores a variável aleatória pode assumir?
   c. A variável aleatória é discreta ou contínua?

Aplicações

AUTOTESTE

3. Três estudantes têm entrevistas programadas no Brookwood Institute com o objetivo de obter empregos de verão. Em cada caso, a entrevista resultará na oferta de um cargo ou em uma recusa. Os resultados experimentais são definidos em termos dos resultados das três entrevistas.
   a. Liste os resultados experimentais.
   b. Defina uma variável aleatória que represente o número de ofertas feitas. A variável é discreta ou contínua?
   c. Mostre o valor da variável aleatória correspondente a cada um dos resultados experimentais.
4. Suponha que saibamos quais são as taxas de hipoteca residencial de 12 instituições de empréstimo da Flórida. Suponha que a variável aleatória de interesse seja o número de instituições de empréstimo pertencentes a esse grupo que oferecem uma taxa fixa de 8,5% ou menos durante 30 anos. Quais valores essa variável aleatória pode assumir?
5. Para realizar certo tipo de análise sangüínea os técnicos de laboratório precisam levar a efeito dois procedimentos. O primeiro procedimento necessita de uma ou duas etapas distintas, e o segundo procedimento requer uma, duas ou três etapas.
   a. Liste os resultados experimentais associados à realização da análise sangüínea.
   b. Se a variável aleatória de interesse for o número total de etapas necessárias para a análise completa (ambos os procedimentos), mostre qual valor a variável aleatória assumirá para cada um dos resultados experimentais.
6. Uma série de experimentos e as variáveis aleatórias correspondentes são listados a seguir. Em cada caso, identifique os valores que a variável aleatória pode assumir e estabeleça se a variável aleatória é discreta ou contínua.

Experimento	Variável aleatória (x)
a. Fazer um exame com 20 questões	Número de questões respondidas corretamente
b. Observar carros que chegam a um posto de pedágio durante uma hora	Número de carros que chegam ao posto de pedágio
c. Fazer a auditoria de 50 declarações de imposto	Número de declarações que contêm erros
d. Observar o trabalho de um empregado	Número de horas não produtivas em um dia de trabalho de oito horas
e. Pesar um carregamento de produtos	Número de quilos

5.2 DISTRIBUIÇÕES DISCRETAS DE PROBABILIDADE

A **distribuição de probabilidade** de uma variável aleatória descreve como as probabilidades estão distribuídas sobre os valores da variável aleatória. Para uma variável discreta x, a distribuição de probabilidade é definida por uma **função probabilidade**, denotada por $f(x)$. A função probabilidade fornece a probabilidade correspondente a cada um dos valores da variável aleatória.

Como ilustração de uma variável aleatória discreta e sua distribuição de probabilidade, considere as vendas de automóveis na DiCarlo Motors, em Saratoga, Nova York. Nos últimos 300 dias de operação, os dados de vendas mostram 54 dias sem vendas de automóveis, 117 dias com um automóvel vendido, 72 dias com dois automóveis vendidos, 42 dias com três automóveis vendidos, 12 dias com quatro automóveis vendidos e três dias com cinco automóveis vendidos. Suponha que consideremos o experimento de selecionar um dia de operação na DiCarlo Motors. Definimos a variável aleatória de interesse como x = o número de automóveis vendidos durante um dia. A partir de dados históricos, sabemos que x é uma variável aleatória discreta que pode assumir os valores 0, 1, 2, 3, 4 ou 5. Na notação da função probabilidade, $f(0)$ fornece a probabilidade de 0 automóveis vendidos, $f(1)$ fornece a probabilidade de um automóvel vendido e assim por diante. Uma vez que os dados históricos mostram 54 dos 300 dias com 0, atribuímos o valor $54/300 = 0{,}18$ para $f(0)$, indicando que a probabilidade de 0 automóvel ter sido vendido durante um dia é de 0,18. Analogamente, uma vez que 117 de 300 dias tiveram um automóvel vendido, atribuímos o valor de $17/300 = 0{,}39$ para $f(1)$, indicando que a probabilidade de exatamente um automóvel ter sido vendido durante um dia é de 0,39. Continuando desse modo para outros valores da variável aleatória, calculamos os valores para $f(2)$, $f(3)$, $f(4)$ e $f(5)$, como mostra a Tabela 5.3, a distribuição de probabilidade para o número de automóveis vendidos durante um dia na DiCarlo Motors.

A principal vantagem de definir uma variável aleatória e sua distribuição de probabilidade é que, uma vez que a distribuição de probabilidade seja conhecida, torna-se relativamente fácil determinar a probabilidade de uma série de eventos que podem ser do interesse de um tomador de decisões. Por exemplo, usando a distribuição de probabilidade na DiCarlo Motors, como mostrado na Tabela 5.3, vemos que o número mais provável de automóveis vendidos durante um dia é 1, com a probabilidade de $f(1) = 0{,}39$. Além disso, há uma probabilidade $f(3) + f(4) + f(5) = 0{,}14 + 0{,}04 + 0{,}01 = 0{,}19$ de venderem três automóveis ou mais durante um dia. Essas probabilidades, além de outras que um tomador de decisões pode solicitar, fornecem a informação que pode auxiliá-lo a entender o processo de venda de automóveis na DiCarlo Motors.

No desenvolvimento de uma função probabilidade para qualquer variável discreta, as duas condições seguintes precisam ser satisfeitas.

Essas condições são análogas às duas exigências básicas para atribuir probabilidades aos resultados experimentais apresentados no Capítulo 4.

CONDIÇÕES NECESSÁRIAS PARA UMA FUNÇÃO PROBABILIDADE DISCRETA

$$f(x) \geq 0 \quad \textbf{(5.1)}$$

$$\Sigma f(x) = 1 \quad \textbf{(5.2)}$$

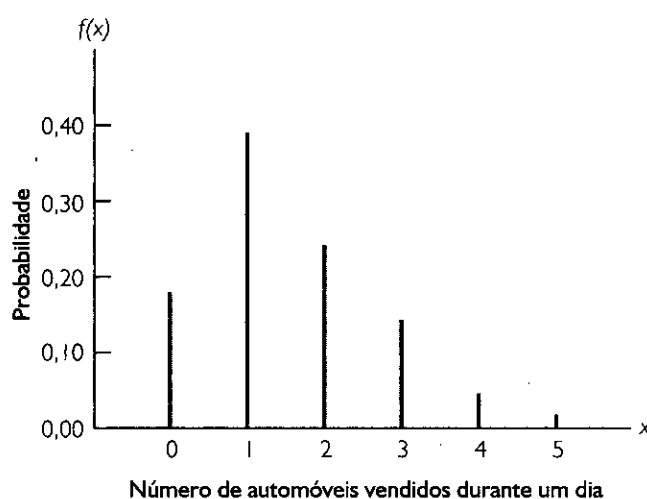
A Tabela 5.3 mostra que as probabilidades correspondentes à variável aleatória x satisfazem a condição da Equação 5.1; $f(x)$ é maior ou igual a 0 para todos os valores de x. Além disso, as probabilidades somam 1, de modo que a Equação 5.2 está satisfeita. Assim, a função probabilidade da DiCarlo Motors é uma função probabilidade discreta válida.

Podemos também apresentar graficamente as distribuições de probabilidade. Na Figura 5.1, os valores da variável aleatória x para a DiCarlo Motors são mostrados no eixo horizontal e a probabilidade associada a esses valores é mostrada no eixo vertical.

Além de tabelas e gráficos, freqüentemente se usa uma expressão matemática para descrever as distribuições de probabilidade, a qual fornece a função probabilidade $f(x)$ para cada valor de x. O exemplo mais simples de distribuição de probabilidade discreta apresentado por meio de uma expressão matemática é a **distribuição uniforme de probabilidade discreta**. Sua função probabilidade é definida pela Equação 5.3.

Tabela 5.3 Distribuição de probabilidade correspondente ao número de automóveis vendidos durante um dia na DiCarlo Motors

x	$f(x)$
0	0,18
1	0,39
2	0,24
3	0,14
4	0,04
5	0,01
Total	1,00

Figura 5.1 Representação gráfica da distribuição de probabilidade para o número de automóveis vendidos durante um dia na DiCarlo Motors

FUNÇÃO PROBABILIDADE DISCRETA UNIFORME

$$f(x) = 1/n \tag{5.3}$$

em que:

n = o número de valores que a variável aleatória pode assumir

Por exemplo, considere o experimento de lançar um dado e defina a variável aleatória x como o número que vai surgir. Existem $n = 6$ valores possíveis para a variável aleatória; $x = 1, 2, 3, 4, 5, 6$. Assim, a função probabilidade para essa variável aleatória discreta uniforme é

$$f(x) = 1/6 \qquad x = 1, 2, 3, 4, 5, 6$$

Os valores possíveis da variável aleatória e as probabilidades correspondentes são mostrados a seguir.

x	$f(x)$
1	1/6
2	1/6
3	1/6
4	1/6
5	1/6
6	1/6

Como outro exemplo, considere a variável aleatória x com a seguinte distribuição de probabilidade discreta.

x	$f(x)$
1	1/10
2	2/10
3	3/10
4	4/10

Essa distribuição de probabilidade pode ser definida pela expressão matemática:

$$f(x) = \frac{x}{10} \quad \text{para } x = 1, 2, 3, \text{ ou } 4$$

Calcular $f(x)$ para determinado valor da variável aleatória fornecerá a probabilidade correspondente. Por exemplo, usando a função probabilidade anterior, vemos que $f(2) = 2/10$ fornece a probabilidade de que a variável aleatória assumirá um valor igual a 2.

As distribuições de probabilidade discretas mais amplamente usadas são, de maneira geral, especificadas por expressões matemáticas. Três casos importantes são a distribuição binomial, a distribuição de Poisson e a distribuição hipergeométrica, todas discutidas posteriormente neste capítulo.

Exercícios

Métodos

AUTOTESTE

7. Segue-se a distribuição de probabilidade da variável aleatória x.

x	$f(x)$
20	0,20
25	0,15
30	0,25
35	0,40

a. Essa distribuição de probabilidade é válida? Explique.
b. Qual é a probabilidade de x ser igual a 30?
c. Qual é a probabilidade de x ser menor ou igual a 25?
d. Qual é a probabilidade de x ser maior que 30?

Aplicações

AUTOTESTE

8. Os dados a seguir foram coletados contando-se o número de salas de cirurgia em uso no Hospital Geral de Tampa em um período de 20 dias: em três dos dias somente uma sala de cirurgia foi usada, em cinco dos dias duas foram usadas, em oito dos dias três foram usadas e em quatro dias todas as quatro salas de cirurgia do hospital foram usadas.

a. Use a abordagem da freqüência relativa para construir a distribuição de probabilidade correspondente ao número de salas de cirurgia em uso em qualquer dia do período.
b. Desenhe um gráfico da distribuição de probabilidade.
c. Mostre que sua distribuição de probabilidade satisfaz as condições necessárias a uma distribuição de probabilidade discreta válida.

9. Nacionalmente, 38% dos estudantes da quarta série do ensino fundamental não conseguem ler um livro apropriado à sua faixa etária. Os dados a seguir mostram o número de crianças, por idade, identificadas como estudantes com dificuldade de aprendizagem sob educação especial. A maioria dessas crianças tem problemas de leitura que devem ser identificados e corrigidos antes da terceira série. A legislação federal vigente nos Estados Unidos não permite que a maioria das crianças receba apoio extra de programas de educação especial até que elas se atrasem aproximadamente dois anos na capacidade de aprendizagem, e isso, tipicamente, significa a terceira série ou mais tarde (*USA Today*, 6 de setembro de 2001).

Idade	Número de Crianças
6	37.369
7	87.436
8	160.840
9	239.719
10	286.719
11	306.533
12	310.787
13	302.604
14	289.168

Suponha que queiramos selecionar uma amostra de crianças identificadas como estudantes com dificuldade de aprendizagem sob educação especial para um programa idealizado para melhorar a capacidade de leitura. Seja x uma variável aleatória que indica a idade de uma criança selecionada aleatoriamente.

a. Use os dados para desenvolver uma distribuição de probabilidade para x. Especifique os valores para a variável aleatória e os valores correspondentes para a função probabilidade $f(x)$.
b. Desenhe um gráfico da distribuição de probabilidade.
c. Mostre que a distribuição de probabilidade satisfaz as Equações (5.1) e (5.2).

10. A Tabela 5.4 mostra as distribuições de freqüência percentuais das pontuações de satisfação no trabalho referentes a uma amostra de executivos seniores de sistemas de informação e gerentes de nível médio de sistemas de informação. As pontuações variam do baixo valor 1 (muito insatisfeitos) ao elevado valor 5 (muito satisfeitos).

Tabela 5.4 Distribuição de freqüência percentual das pontuações de satisfação no trabalho referentes a executivos e gerentes de nível médio de sistemas de informação

Pontuação de Satisfação no Trabalho	Executivos Seniores de Sistemas de Informação (%)	Gerentes de Nível Médio de Sistemas de Informação (%)
1	5	4
2	9	10
3	3	12
4	42	46
5	41	28

a. Desenvolva uma distribuição de probabilidade referente à pontuação da satisfação de um executivo sênior no trabalho.
b. Desenvolva a distribuição de probabilidade referente à pontuação da satisfação de um gerente médio no trabalho.
c. Qual é a probabilidade de um executivo sênior registrar uma pontuação de satisfação no trabalho igual a 4 ou 5?
d. Qual é a probabilidade de um gerente de nível médio estar muito satisfeito?
e. Compare a satisfação global no trabalho dos executivos seniores e dos gerentes de nível médio.

11. Um técnico faz manutenção de máquinas de postagem em empresas na região de Phoenix. Dependendo do tipo de defeito, uma visita técnica pode demandar 1, 2, 3 ou 4 horas. Os diferentes tipos de defeito ocorrem aproximadamente na mesma freqüência.

a. Desenvolva uma distribuição de probabilidade para a duração de uma visita técnica.
b. Desenhe um gráfico da distribuição de probabilidade.
c. Mostre que sua distribuição de probabilidade satisfaz as condições necessárias a uma função probabilidade discreta.
d. Qual é a probabilidade de a visita técnica demandar três horas?
e. Uma visita técnica acabou de chegar, mas o tipo de defeito é desconhecido. São 15h e o técnico habitualmente deixa o trabalho às 17h. Qual é a probabilidade de o técnico precisar trabalhar em hora extra para consertar a máquina ainda hoje?

12. O diretor de admissão do Lakeville Community College avaliou subjetivamente uma distribuição de probabilidade para x, equivalente ao número de matriculandos, da seguinte maneira:

x	$f(x)$
1.000	0,15
1.100	0,20
1.200	0,30
1.300	0,25
1.400	0,10

a. Essa é uma distribuição de probabilidade válida? Explique.
b. Qual é a probabilidade de 1.200 estudantes ou menos se matricularem?

13. Um psicólogo determinou que o número de sessões necessárias para conquistar a confiança de um novo paciente pode ser de 1, 2 ou 3. Seja x uma variável aleatória que indica o número de sessões necessárias para conquistar a confiança do paciente. A seguinte função de probabilidade foi proposta.

$$f(x) = \frac{x}{6} \quad \text{para } x = 1, 2, \text{ ou } 3$$

a. Essa é uma função probabilidade válida? Explique.
b. Qual é a probabilidade de serem necessárias exatamente duas sessões para conquistar a confiança do paciente?
c. Qual é a probabilidade de serem necessárias pelo menos duas sessões para conquistar a confiança do paciente?

14. A tabela seguinte é uma distribuição de probabilidade parcial referente ao lucro projetado da MRA Company (x = lucro em milhares de dólares) para o primeiro ano de operação (o valor negativo denota um prejuízo).

x	$f(x)$
−100	0,10
0	0,20
50	0,30
100	0,25
150	0,10
200	

a. Qual é o valor adequado para $f(200)$? Qual é a sua interpretação desse valor?
b. Qual é a probabilidade de a MRA ser rentável?
c. Qual é a probabilidade de a MRA alcançar pelo menos US$ 100 mil?

5.3 VALOR ESPERADO E VARIÂNCIA

Valor Esperado

O **valor esperado**, ou média, de uma variável aleatória é a medida da posição central da variável aleatória. A expressão matemática do valor esperado para a variável aleatória discreta x é dada a seguir.

O valor esperado é a média ponderada dos valores que a variável aleatória pode assumir. Os pesos são as probabilidades.

VALOR ESPERADO DE UMA VARIÁVEL ALEATÓRIA DISCRETA

$$E(x) = \mu = \Sigma x f(x) \qquad \textbf{(5.4)}$$

Tanto a notação $E(x)$ como μ podem ser usadas para denotar o valor esperado de uma variável aleatória.

O valor esperado não precisa ser um valor que a variável aleatória possa assumir.

A Equação 5.4 mostra que para calcular o valor esperado de uma variável aleatória discreta precisamos multiplicar cada um dos valores da variável aleatória pela probabilidade $f(x)$ correspondente e, então, adicionar os produtos resultantes. Usando o exemplo das vendas de automóveis da DiCarlo Motors da Seção 5.2, mostramos na Tabela 5.5 os cálculos do valor esperado referentes ao número de automóveis vendidos durante um dia. A soma das entradas na coluna $xf(x)$ mostra que o valor esperado é de 1,50 automóvel por dia. Sabemos, portanto, que, embora seja possível a realização de 0, 1, 2, 3, 4 ou 5 vendas de automóveis em qualquer um dos dias, ao longo do tempo a DiCarlo pode prever a venda de uma média de 1,50 automóvel por dia. Supondo 30 dias de operação durante um mês, podemos usar o valor esperado de 1,50 para prever vendas mensais médias de 30(1,50) = 45 automóveis.

Variância

A variância é a média ponderada dos desvios elevados ao quadrado que uma variável aleatória sofre a partir de sua média. Os pesos são as probabilidades.

Não obstante o valor esperado fornecer o valor médio para a variável aleatória, freqüentemente necessitamos de uma medida de variabilidade, ou de dispersão. Tal como usamos a variância no Capítulo 3 para sintetizar a variabilidade no conjunto de dados, usamos agora a **variância** para sintetizar a variabilidade nos valores da variável aleatória. A expressão matemática para a variância de variável aleatória discreta é apresentada a seguir.

VARIÂNCIA DE UMA VARIÁVEL ALEATÓRIA DISCRETA

$$\text{Var}(x) = \sigma^2 = \Sigma(x - \mu)^2 f(x) \qquad (5.5)$$

Tabela 5.5 Cálculo do valor esperado para o número de automóveis vendidos durante um dia na DiCarlo Motors

x	$f(x)$	$xf(x)$
0	0,18	0(0,18) = 0,00
1	0,39	1(0,39) = 0,39
2	0,24	2(0,24) = 0,48
3	0,14	3(0,14) = 0,42
4	0,04	4(0,04) = 0,16
5	0,01	5(0,01) = 0,05
	$E(x) = \mu = \Sigma xf(x)$	1,50

Tabela 5.6 Cálculo da variância para o número de automóveis vendidos durante um dia na Dicarlo Motors

x	$x - \mu$	$(x - \mu)^2$	$f(x)$	$(x - \mu)^2 f(x)$
0	0 − 1,50 = −1,50	2,25	0,18	2,25(0,18) = 0,4050
1	1 − 1,50 = −0,50	0,25	0,39	0,25(0,39) = 0,0975
2	2 − 1,50 = 0,50	0,25	0,24	0,25(0,24) = 0,0600
3	3 − 1,50 = 1,50	2,25	0,14	2,25(0,14) = 0,3150
4	4 − 1,50 = 2,50	6,25	0,04	6,25(0,04) = 0,2500
5	5 − 1,50 = 3,50	12,25	0,01	12,25(0,01) = 0,1225
			$\sigma^2 = \Sigma(x - \mu)^2 f(x)$	1,2500

Como mostra a Equação 5.5, uma parte fundamental da fórmula da variância é o desvio, $x - \mu$, que mede quão distante um valor em particular da variável aleatória se encontra do valor esperado, ou média, μ. No cálculo da variância de uma variável aleatória, os desvios são elevados ao quadrado e então ponderados pelo valor correspondente da função probabilidade. A soma desses desvios elevados ao quadrado ponderados para todos os valores da variável aleatória denomina-se *variância*. As notações Var(x) e σ^2 são ambas utilizadas para denotar a variância de uma variável aleatória.

O cálculo da variância para a distribuição de probabilidade do número de automóveis vendidos durante um dia na DiCarlo Motors está resumido na Tabela 5.6. Notamos que a variância é 1,25. **O desvio padrão,** σ, é definido como a raiz quadrada positiva da variância. Assim, o desvio padrão do número de automóveis vendidos durante um dia é

$$\sigma = \sqrt{1,25} = 1,118$$

O desvio padrão é medido nas mesmas unidades que a variável aleatória (σ = 1,118 automóvel) e, portanto, freqüentemente é preferido para descrever a variabilidade de uma variável aleatória. A variância σ^2 é medida em unidades elevadas ao quadrado e, desse modo, é mais difícil de ser interpretada.

Exercícios

Métodos

15. A tabela seguinte apresenta uma distribuição de probabilidade referente à variável aleatória x.

x	$f(x)$
3	0,25
6	0,50
9	0,25

a. Calcule $E(x)$, o valor esperado de x.
b. Calcule σ^2, a variância de x.
c. Calcule σ, o desvio padrão de x.

16. A tabela seguinte apresenta uma distribuição de probabilidade referente à variável aleatória y.

AUTOTESTE

x	$f(y)$
3	0,25
4	0,30
7	0,40
8	0,10

a. Calcule $E(y)$.
b. Calcule Var(y) e σ.

Aplicações

17. Um serviço voluntário de ambulâncias atende de 0 a 5 chamadas de serviço em determinado dia. A distribuição de probabilidade correspondente ao número de chamadas de serviço é apresentada a seguir.

Número de Chamadas de Serviço	Probabilidade	Número de Chamadas de Serviço	Probabilidade
0	0,10	3	0,20
1	0,15	4	0,15
2	0,30	5	0,10

a. Qual é o número esperado de chamadas de serviço?
b. Qual é a variância no número de chamadas de serviço? Qual é o desvio padrão?

18. A American Housing Survey registrou os seguintes dados sobre o número de quartos de dormir em casas ocupadas por proprietários e casas ocupadas por locatários em grandes cidades (http://www.census.gov, 31 de março de 2003).

AUTOTESTE

	Número de Casas (milhares)	
Quartos de Dormir	Ocupadas por Locatários	Ocupadas por Proprietários
0	547	23
1	5.012	541
2	6.100	3.832
3	2.644	8.690
4 ou mais	557	3.783

a. Defina uma variável aleatória x = o número de quartos de dormir em casas ocupadas por locatários e desenvolva uma distribuição de probabilidade para a variável aleatória. (Digamos que $x = 4$ represente 4 ou mais quartos de dormir.)
b. Calcule o valor esperado e a variância do número de quartos de dormir em casas ocupadas por locatários.
c. Defina uma variável aleatória y = o número de quartos de dormir em casas ocupadas por proprietários e desenvolva uma distribuição de probabilidade para a variável aleatória. (Digamos que y = 4 represente 4 ou mais quartos de dormir.)
d. Calcule o valor esperado e a variância do número de quartos de dormir em casas ocupadas por proprietários.
e. Quais observações você é capaz de fazer a partir de uma comparação do número de quartos de dormir em casas ocupadas por locatários e casas ocupadas por proprietários?

19. A National Basketball Association (NBA) registra uma série de estatísticas para cada time. Duas dessas estatísticas são a porcentagem de *field goals*[1] e a porcentagem de lances de três pontos realizados pelo time. Durante uma parte da temporada de 2004, os registros de lances dos 29 times da NBA mostravam que a probabilidade de marcarem dois pontos fazendo um *field goal* era 0,44, e que a probabilidade de marcarem três pontos fazendo um lance de três pontos era 0,34 (http://www.nba.com, 3 de janeiro de 2004).
   a. Qual é o valor esperado de um arremesso de dois pontos para esses times?
   b. Qual é o valor esperado de um arremesso de três pontos para esses times?
   c. Se a probabilidade de fazer um arremesso de dois pontos é maior que a probabilidade de fazer um arremesso de três pontos, por que os técnicos permitem a alguns jogadores fazerem arremessos de três pontos quando têm a oportunidade? Use o valor esperado para explicar sua resposta.
20. A distribuição de probabilidade para reclamação de danos sobre seguros de colisão pagos pela Newton Automobile Insurance Company é mostrada a seguir.

Pagamento ($)	Probabilidade
0	0,90
400	0,04
1.000	0,03
2.000	0,01
4.000	0,01
6.000	0,01

   a. Use o pagamento de colisão esperado para determinar o prêmio de seguro de colisão que possibilitaria à empresa não ter lucro nem prejuízo.
   b. A companhia de seguros cobra uma taxa anual de US$ 260 para a cobertura de colisão. Qual é o valor esperado da apólice de seguro contra colisão para o proprietário da apólice? (*Dica:* Esse valor é o pagamento esperado da companhia menos o custo de cobertura.) Por que o proprietário da apólice compra uma apólice de colisão com esse valor esperado?
21. As seguintes pontuações de satisfação no trabalho referentes a uma amostra de executivos seniores de sistemas de informação e de gerentes de nível médio de sistemas de informação variam do baixo valor 1 (muito insatisfeitos) ao elevado valor 5 (muito satisfeitos).

	Probabilidade	
Pontuação de Satisfação no Trabalho	Executivos Seniores de Sistemas de Informação	Gerentes de Nível Médio de Sistemas de Informação
1	0,05	0,04
2	0,09	0,10
3	0,03	0,12
4	0,42	0,46
5	0,41	0,28

   a. Qual é o valor esperado da pontuação de satisfação no trabalho para os executivos seniores?
   b. Qual é o valor esperado da pontuação de satisfação no trabalho para os gerentes de nível médio?
   c. Calcule a variância das pontuações de satisfação no trabalho para os executivos e os gerentes de nível médio.
   d. Calcule o desvio padrão das pontuações de satisfação no trabalho para ambas as distribuições de probabilidade.
   e. Calcule a satisfação global no trabalho dos executivos seniores e dos gerentes de nível médio.
22. A demanda por um produto da Carolina Industries varia muito de mês a mês. A distribuição de probabilidade na tabela a seguir, baseada nos dados dos últimos dois anos, mostra a demanda mensal da empresa.

Demanda Unitária	Probabilidade
300	0,20
400	0,30
500	0,35
600	0,15

[1] NT: *Field goal*: Arremesso que marca dois pontos, e se for de certa distância, três pontos (basquete).

a. Se a empresa basear os pedidos de compra mensais no valor esperado da demanda mensal, qual deve ser o lote de compra mensal da Carolina Industries para esse produto?
b. Considere que cada unidade demandada gera US$ 70 de receita e que cada unidade encomendada custa US$ 50. Quanto a empresa ganhará ou perderá em um mês se vier a colocar um pedido de compra baseando-se em sua resposta ao item (a) e se a demanda real pelo item for de 300 unidades?

23. A 2002 New York City Housing and Vacancy Survey mostrou um total de 59.324 unidades residenciais *rent-controlled*[2] e 236.263 unidades *rent-stabilized*[3] construídas em 1947 ou mais tarde. Para essas unidades de aluguel, as distribuições de probabilidade referentes ao número de pessoas que moram na unidade são apresentadas a seguir (http://www.census.gov, 12 de janeiro de 2004).

Número de pessoas	*Rent-Controlled*	*Rent-Stabilized*
1	0, 61	0,41
2	0,27	0,30
3	0,07	0,14
4	0,04	0,11
5	0,01	0,03
6	0,00	0,01

a. Qual é o valor esperado do número de pessoas que moram em cada tipo de unidade?
b. Qual é a variância do número de pessoas que moram em cada tipo de unidade?
c. Faça algumas comparações entre o número de pessoas que moram em unidades *rent-controlled* e o número de pessoas que moram em unidades *rent-stabilized.*

24. A J. R. Ryland Computer Company está considerando uma expansão de fábrica que tornará possível à empresa começar a produzir um novo tipo de computador. O presidente da empresa precisa determinar se faz a expansão em média ou em grande escala. Uma incerteza é a demanda do novo produto, a qual, para propósitos de planejamento, pode ter uma baixa demanda, uma média demanda ou uma alta demanda. As estimativas de probabilidades de demandas são 0,20; 0,50; e 0,30, respectivamente. Se x e y indicam o lucro anual em milhares de dólares, os planejadores da empresa desenvolveram as seguintes previsões de lucro para os projetos de expansão de média e de grande escalas.

		Lucro da Expansão de Média Escala		Lucro da Expansão de Grande Escala	
		x	$f(x)$	y	$f(y)$
	Baixa	50	0,20	0	0,20
Demanda	Média	150	0,50	100	0,50
	Elevada	200	0,30	300	0,30

a. Calcule o valor esperado para o lucro associado às duas alternativas de expansão. Qual decisão é preferível para o objetivo de maximizar o lucro esperado?
b. Calcule a variância para o lucro associado às duas alternativas de expansão. Qual decisão é preferível para o objetivo de minimizar o risco ou a incerteza?

5.4 DISTRIBUIÇÃO DE PROBABILIDADE BINOMIAL

A distribuição de probabilidade binomial é uma distribuição de probabilidade discreta que tem muitas aplicações. Ela está associada a um experimento de múltiplas etapas que chamamos experimento binomial.

[2] NT: *Rent-controlled apartment*: Para um apartamento ser *rent-controlled*, o inquilino deve residir nele continuamente desde 1º de julho de 1974. Quando um apartamento *rent-controlled* é desocupado, ele se torna automaticamente *rent-stabilized* ou sua regulamentação é cancelada (Estados Unidos).

[3] NT: *Rent-stabilized apartment*: Unidades *rent-stabilized* são aqueles apartamentos em prédios de seis ou mais unidades construídos entre 1º de fevereiro de 1947 e 1º de janeiro de 1974. Os inquilinos têm o direito de receber os serviços necessários, ter a renovação de seus contratos de aluguel e não podem ser despejados a não ser nos termos da lei (Estados Unidos).

Um Experimento Binomial

Um **experimento binomial** tem as quatro propriedades seguintes:

PROPRIEDADES DE UM EXPERIMENTO BINOMIAL

1. O experimento consiste em uma seqüência de n ensaios idênticos.
2. Dois resultados são possíveis em cada ensaio. Referimo-nos a um como um *sucesso* e ao outro como um *fracasso*.
3. A probabilidade de um sucesso, denotado por p, não se modifica de ensaio para ensaio. Conseqüentemente, a probabilidade de um fracasso, denotado por $1 - p$, não se modifica de ensaio para ensaio.
4. Os ensaios são independentes.

Jakob Bernoulli (1654-1705), o primeiro Bernoulli da família de matemáticos suíços, publicou um tratado das probabilidades que continha a teoria das permutações e combinações, bem como o teorema binomial.

Se as propriedades 2, 3 e 4 estão presentes, dizemos que os ensaios são gerados por um processo de Bernoulli. Se, além disso, a propriedade 1 está presente, dizemos que temos um experimento binomial. A Figura 5.2 retrata uma seqüência possível de sucessos e fracassos de um experimento binomial envolvendo oito ensaios.

Em um experimento binomial, nosso interesse é o *número de sucessos que ocorrem nos* n *ensaios*. Se x denota o número de sucessos que ocorrem nos n ensaios, vemos que x pode assumir os valores de 0, 1, 2, 3, ..., n. Uma vez que o número de valores é finito, x é uma variável aleatória *discreta*. A distribuição de probabilidade associada a essa variável aleatória é chamada de **distribuição de probabilidade binomial**. Por exemplo, considere o experimento de jogar uma moeda cinco vezes e em cada arremesso observar se a moeda cai com coroa ou com cara voltada para cima. Suponha que estejamos interessados em contar o número de caras que aparecem nos cinco arremessos. Esse experimento tem as propriedades de um experimento binomial? Qual é a variável aleatória de interesse? Observe que:

1. O experimento consiste em cinco ensaios idênticos; cada ensaio envolve o lançamento de uma moeda.
2. Dois resultados são possíveis para cada ensaio: uma cara ou uma coroa. Podemos designar cara um sucesso e coroa um fracasso.
3. A probabilidade de se obter cara e a probabilidade de se obter coroa são as mesmas para cada ensaio, com $p = 0,5$ e $1 - p = 0,5$.
4. Os ensaios ou arremessos são independentes porque o resultado de qualquer um dos ensaios não é afetado pelo que acontece nos outros ensaios ou arremessos.

Figura 5.2 Uma seqüência possível de sucessos e fracassos para um experimento binomial de oito ensaios

Propriedade 1: O experimento consiste em n = 8 ensaios.

Propriedade 2: Cada ensaio resulta em sucesso (S) ou fracasso (F).

Ensaios →	1	2	3	4	5	6	7	8
Resultados →	S	F	F	S	S	F	S	S

Desse modo, as propriedades de um experimento binomial estão satisfeitas. A variável aleatória de interesse é x = o número de caras que aparece nos cinco ensaios. Nesse caso, x pode assumir os valores 0, 1, 2, 3, 4 ou 5.

Como outro exemplo, considere um vendedor de seguros que visita dez famílias selecionadas aleatoriamente. O resultado associado a cada visita é classificado como um sucesso se a família comprar uma apólice de seguros, e como um fracasso se a família não comprar. Por experiência, o vendedor sabe que a probabilidade de uma família selecionada aleatoriamente comprar uma apólice de seguro é igual a 0,10. Verificando as propriedades de um experimento binomial, observamos que:

1. O experimento consiste em dez ensaios idênticos; cada ensaio envolve contatar uma família.
2. Dois resultados são possíveis em cada ensaio: a família compra uma apólice (sucesso) ou a família não compra uma apólice (fracasso).
3. Considera-se que as probabilidades de uma compra e de uma não-compra são as mesmas para cada contato de venda, com $p = 0{,}10$ e $1 - p = 0{,}90$.
4. Os ensaios são independentes porque as famílias são selecionadas aleatoriamente.

Como as quatro hipóteses estão satisfeitas, esse exemplo é um experimento binomial. A variável aleatória de interesse é o número de vendas obtidas ao contatar as dez famílias. Nesse caso, x pode assumir os valores 0, 1, 2, 3, 4, 5, 6, 7, 8, 9 e 10.

A propriedade 3 do experimento binomial é chamada *hipótese estacionária,* e é confundida algumas vezes com a propriedade 4, independência dos ensaios. Para ver como elas diferem, considere outra vez o caso do vendedor que contata famílias para vender apólices de seguro. Se, no decorrer do dia, o vendedor se cansar e perder o entusiasmo, a probabilidade de sucesso (vender uma apólice) pode cair para 0,05, por exemplo, lá pela décima ligação. Nesse caso, a propriedade 3 (imutabilidade) não seria satisfeita, e não teríamos um experimento binomial. Mesmo que a propriedade 4 se mantivesse – isto é, as decisões de compra de cada família fossem tomadas independentemente –, não seria um experimento binomial se a propriedade 3 não fosse satisfeita.

Em aplicações que envolvem experimentos binomiais, uma fórmula matemática especial, denominada *função de probabilidade binomial*, pode ser usada para calcular a probabilidade de x sucessos nos n ensaios. Usando os conceitos de probabilidade apresentados no Capítulo 4, mostraremos no contexto de um problema ilustrativo como a fórmula pode ser desenvolvida.

O Problema da Loja de Roupas do Martin

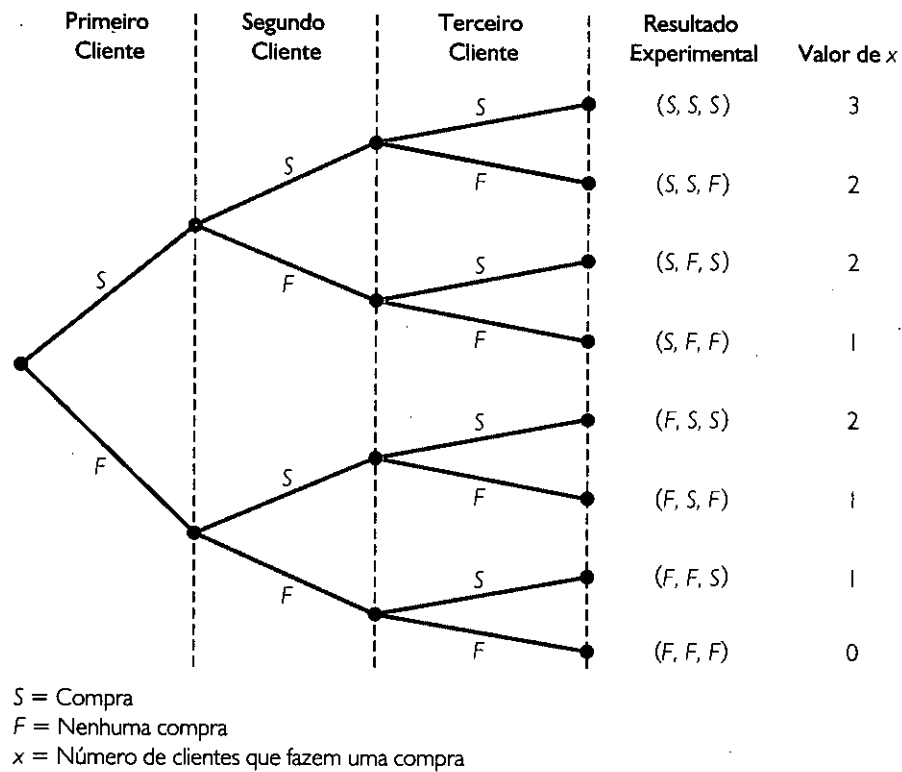
Consideremos as decisões de compra dos próximos três clientes que entram na loja de roupas do Martin. Com base em sua experiência, o gerente da loja estima que a probabilidade de qualquer dos clientes comprar é de 0,30. Qual é a probabilidade de dois dos próximos três clientes realizarem uma compra?

Usando um diagrama em árvore (Figura 5.3), podemos ver que o experimento de observar os três clientes, cada um deles tomando uma decisão de compra, tem oito resultados possíveis. Usando S para denotar sucesso (uma compra) e F para denotar fracasso (nenhuma compra), estamos interessados nos resultados experimentais que envolvem dois sucessos nos três ensaios (decisões de compra). A seguir, vamos verificar que o experimento envolvendo a seqüência de três decisões de compra pode ser visto como um experimento binomial. Verificando as quatro exigências para um experimento binomial, notamos que:

1. O experimento pode ser descrito como uma seqüência de três ensaios idênticos, sendo um ensaio para cada um dos três clientes que entrarão na loja.
2. Dois resultados – o cliente faz uma compra (sucesso) ou o cliente não faz uma compra (fracasso) – são possíveis para cada ensaio.
3. A probabilidade de o cliente vir a fazer uma compra (0,30) ou não fazer uma compra (0,70) é considerada a mesma para todos os clientes.
4. A decisão de compra de cada cliente é independente das decisões de outros clientes.

Figura 5.3 Diagrama em árvore para o problema da loja de roupas do Martin

Primeiro Cliente	Segundo Cliente	Terceiro Cliente	Resultado Experimental	Valor de x
S	S	S	(S, S, S)	3
		F	(S, S, F)	2
	F	S	(S, F, S)	2
		F	(S, F, F)	1
F	S	S	(F, S, S)	2
		F	(F, S, F)	1
	F	S	(F, F, S)	1
		F	(F, F, F)	0

S = Compra
F = Nenhuma compra
x = Número de clientes que fazem uma compra

Portanto, as propriedades de um experimento binomial estão presentes.

O número de resultados experimentais que resultam em exatamente x sucessos em n ensaios pode ser calculado a partir da seguinte fórmula.[4]

NÚMERO DE RESULTADOS EXPERIMENTAIS QUE FORNECEM EXATAMENTE X SUCESSOS EM N ENSAIOS

$$\binom{n}{x} = \frac{n!}{x!(n-x)!} \tag{5.6}$$

em que

$$n! = n(n-1)(n-2)\cdots(2)(1)$$

e, por definição,

$$0! = 1$$

Retornemos agora ao experimento da loja de roupas do Martin, envolvendo as decisões de compra tomadas por três clientes.

A Equação 5.6 pode ser usada para determinar o número de resultados experimentais envolvendo duas compras; isto é, o número de modos de se obter $x = 2$ sucessos nos $n = 3$ ensaios. Da Equação 5.6, temos:

$$\binom{n}{x} = \binom{3}{2} = \frac{3!}{2!(3-2)!} = \frac{(3)(2)(1)}{(2)(1)(1)} = \frac{6}{2} = 3$$

[4] Essa fórmula, apresentada no Capítulo 4, determina o número de combinações de n objetos x selecionados a cada vez. Para o experimento binomial, essa fórmula combinatória fornece o número de resultados experimentais (seqüências de n ensaios) resultantes em x sucessos.

A Equação 5.6 mostra que três dos resultados experimentais produzem dois sucessos. Da Figura 5.3, vemos que esses três resultados são denotados por (S, S, F), (S, F, S) e (F, S, S).

Usando a Equação 5.6 para determinar quantos resultados experimentais obtêm três sucessos (compras) nos três ensaios, obtemos

$$\binom{n}{x} = \binom{3}{3} = \frac{3!}{3!(3-3)!} = \frac{3!}{3!0!} = \frac{(3)(2)(1)}{3(2)(1)(1)} = \frac{6}{6} = 1$$

Da Figura 5.3, vemos que um resultado experimental com três sucessos é identificado por (S, S, S).

Sabemos que a Equação (5.6) pode ser usada para determinar o número de resultados experimentais que resultam em x sucessos. Mas, se quisermos estabelecer a probabilidade de x sucessos em n ensaios, precisamos também conhecer a probabilidade associada a cada um desses resultados experimentais. Uma vez que os ensaios de um experimento binomial são independentes, podemos simplesmente multiplicar as probabilidades associadas a cada resultado experimental para encontrar a probabilidade de uma seqüência de sucessos e fracassos em particular.

A probabilidade de compras efetuadas pelos primeiros dois clientes e de nenhuma compra pelo terceiro cliente, denotada por (S, S, F), é dada por

$$pp(1 - p)$$

Com 0,30 de probabilidade de uma compra em qualquer um dos ensaios, a probabilidade de uma compra nos dois primeiros ensaios e de nenhuma compra no terceiro é dada por

$$(0{,}30)(0{,}30)(0{,}70) = (0{,}30)^2(0{,}70) = 0{,}063$$

Dois outros resultados experimentais também resultam em dois sucessos e um fracasso. As probabilidades referentes a todas as três seqüências envolvendo dois sucessos são mostradas a seguir.

Resultados Experimentais				
Primeiro Cliente	Segundo Cliente	Terceiro Cliente	Resultado Experimental	Probabilidade do Resultado Experimental
Compra	Compra	Nenhuma Compra	(S, S, F)	$pp(1-p) = p^2(1-p)$ $= (0{,}30)^2(0{,}70) = 0{,}063$
Compra	Nenhuma Compra	Compra	(S, F, S)	$p(1-p)p = p^2(1-p)$ $= (0{,}30)^2(0{,}70) = 0{,}063$
Nenhuma Compra	Compra	Compra	(F, S, S)	$(1-p)pp = p^2(1-p)$ $= (0{,}30)^2(0{,}70) = 0{,}063$

Observe que todos os três resultados experimentais com dois sucessos têm exatamente a mesma probabilidade. Essa observação se mantém como regra. Em qualquer experimento binomial todas as seqüências de resultados de ensaio que produzem x sucessos em n ensaios têm a *mesma probabilidade* de ocorrência. A probabilidade de cada seqüência de ensaios produzir x sucessos em n ensaios é apresentada a seguir.

Probabilidade de uma seqüência de resultados de ensaio em particular com x sucessos em n ensaios $= p^x(1-p)^{(n-x)}$ **(5.7)**

Em relação à loja de roupas do Martin, essa fórmula mostra que qualquer resultado experimental com dois sucessos tem a probabilidade $p^2(1-p)^{(3-2)} = p^2(1-p)^1 = (0{,}30)^2(0{,}70)^1 = 0{,}063$.

Como a Equação 5.6 mostra o número de resultados em um experimento binomial com x sucessos e a Equação 5.7 fornece a probabilidade referente a cada seqüência envolvendo x sucessos, combinamos as Equações 5.6 e 5.7 para obter a seguinte **função probabilidade binomial**.

FUNÇÃO PROBABILIDADE BINOMIAL

$$f(x) = \binom{n}{x} p^x (1 - p)^{(n-x)} \tag{5.8}$$

em que

$f(x)$ = a probabilidade de x sucessos em n ensaios
n = o número de ensaios
$\binom{n}{x} = \dfrac{n!}{x!(n-x)!}$
p = a probabilidade de sucesso em qualquer dos ensaios
$1 - p$ = a probabilidade de um fracasso em qualquer dos ensaios

No exemplo da loja de roupas do Martin, vamos calcular a probabilidade de nenhum cliente fazer uma compra, exatamente um cliente fazer uma compra, exatamente dois clientes fazerem uma compra e todos os três clientes fazerem uma compra. Os cálculos estão sintetizados na Tabela 5.7, a qual fornece a distribuição de probabilidade do número de clientes que fazem uma compra. A Figura 5.4 corresponde a um gráfico dessa distribuição de probabilidade.

A função probabilidade binomial pode ser aplicada a *qualquer* experimento binomial. Se estamos convencidos de que uma situação exibe as propriedades de um experimento binomial, e se conhecemos os valores de n e p, podemos usar a Equação 5.8 para calcular a probabilidade de x sucessos nos n ensaios.

Tabela 5.7 Distribuição de probabilidade para o número de clientes que fazem uma compra

x	$f(x)$
0	$\frac{3!}{0!3!}(0,30)^0(0,70)^3 = 0,343$
1	$\frac{3!}{1!2!}(0,30)^1(0,70)^2 = 0,441$
2	$\frac{3!}{2!1!}(0,30)^2(0,70)^1 = 0,189$
3	$\frac{3!}{3!0!}(0,30)^3(0,70)^0 = 0,027$
	1,000

Se considerarmos variações no experimento da loja de roupas do Martin, como dez clientes entrando na loja em vez de três, a função de probabilidade binomial dada pela Equação 5.8 ainda é aplicável. Suponha termos um experimento binomial com $n = 10$, $x = 4$ e $p = 0,30$. A probabilidade de realizarmos exatamente quatro vendas para dez clientes que entram na loja é

$$f(4) = \frac{10!}{4!6!}(0,30)^4(0,70)^6 = 0,2001$$

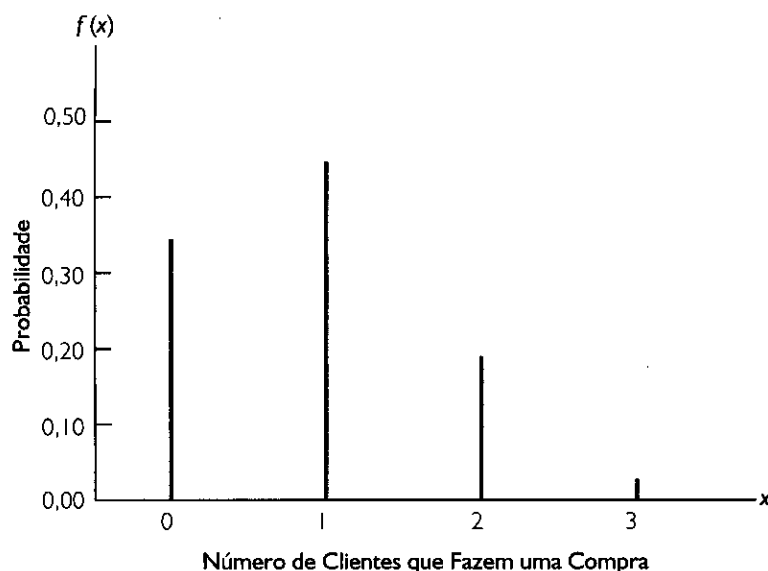
Figura 5.4 Representação gráfica da distribuição de probabilidade para o número de clientes que fazem uma compra

$f(x)$

Probabilidade

0,50
0,40
0,30
0,20
0,10
0,00

0 1 2 3 x

Número de Clientes que Fazem uma Compra

Usando Tabelas de Probabilidades Binomiais

Foram desenvolvidas tabelas que dão a probabilidade de x sucessos em n ensaios para um experimento binomial. Geralmente essas tabelas são fáceis de usar e mais rápidas do que a Equação 5.8. A Tabela 5 do Apêndice B constitui uma dessas tabelas de probabilidades binomiais. Uma parte dessa tabela é apresentada na Tabela 5.8. Para usar essa tabela, precisamos especificar os valores de n, p e x do experimento binomial de interesse. No exemplo apresentado anteriormente da Tabela 5.8, notamos que a probabilidade de $x = 3$ sucessos em um experimento binomial com $n = 10$ e $p = 0{,}40$ é igual a 0,2150. Você pode usar a Equação 5.8 para verificar se viria a obter a mesma resposta se usasse a função de probabilidade binomial diretamente.

Com as modernas calculadoras, essas tabelas são praticamente desnecessárias. É fácil calcular a Equação 5.8 diretamente.

Vamos agora usar a Tabela 5.8 para verificar a probabilidade de quatro sucessos em dez ensaios para o problema da loja de roupas do Martin. Observe que o valor de $f(4) = 0{,}2001$ pode ser lido diretamente na tabela de probabilidades binomiais, com $n = 10$, $x = 4$ e $p = 0{,}30$.

Não obstante as tabelas de probabilidades binomiais serem relativamente fáceis de usar, é impossível haver tabelas que mostrem todos os valores possíveis de n e de p que possam ser encontrados em um experimento binomial. No entanto, com as calculadoras modernas, não é difícil usar a Equação 5.8 para calcular a probabilidade desejada, especialmente se o número de ensaios não for grande. Nos exercícios, você deve praticar o uso da Equação 5.8 para calcular as probabilidades binomiais, a menos que o problema solicite especificamente que você use a tabela de probabilidade binomial.

Pacotes de software de estatística, tais como o Minitab, e pacotes de planilhas eletrônicas, como o Excel, também oferecem a capacidade de calcular probabilidades binomiais. Considere o exemplo da loja de roupas do Martin com $n = 10$ e $p = 0{,}30$. A Figura 5.5 exibe as probabilidades binomiais geradas pelo Minitab para todos os valores possíveis de x. Observe que esses valores são os mesmos que aqueles encontrados na coluna $p = 0{,}30$ da Tabela 5.8. O Apêndice 5.1 apresenta um procedimento passo a passo do uso do Minitab para gerar o resultado apresentado na Figura 5.5. O Apêndice 5.2 descreve como o Excel pode ser usado para calcular probabilidades binomiais.

Tabela 5.8 Valores selecionados da tabela de probabilidades binomiais
Exemplo: $n = 10$, $x = 3$, $p = 0{,}40$; $f(3) = 0{,}2150$

						p					
n	*x*	**0,05**	**0,10**	**0,15**	**0,20**	**0,25**	**0,30**	**0,35**	**0,40**	**0,45**	**0,50**
9	0	0,6302	0,3874	0,2316	0,1342	0,0751	0,0404	0,0207	0,0101	0,0046	0,0020
	1	0,2985	0,3874	0,3679	0,3020	0,2253	0,1556	0,1004	0,0605	0,0339	0,0176
	2	0,0629	0,1722	0,2597	0,3020	0,3003	0,2668	0,2162	0,1612	0,1110	0,0703
	3	0,0077	0,0446	0,1069	0,1762	0,2336	0,2668	0,2716	0,2508	0,2119	0,1641
	4	0,0006	0,0074	0,0283	0,0661	0,1168	0,1715	0,2194	0,2508	0,2600	0,2461
	5	0,0000	0,0008	0,0050	0,0165	0,0389	0,0735	0,1181	0,1672	0,2128	0,2461
	6	0,0000	0,0001	0,0006	0,0028	0,0087	0,0210	0,0424	0,0743	0,1160	0,1641
	7	0,0000	0,0000	0,0000	0,0003	0,0012	0,0039	0,0098	0,0212	0,0407	0,0703
	8	0,0000	0,0000	0,0000	0,0000	0,0001	0,0004	0,0013	0,0035	0,0083	0,0176
	9	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001	0,0003	0,0008	0,0020
10	0	0,5987	0,3487	0,1969	0,1074	0,0563	0,0282	0,0135	0,0060	0,0025	0,0010
	1	0,3151	0,3874	0,3474	0,2684	0,1877	0,1211	0,0725	0,0403	0,0207	0,0098
	2	0,0746	0,1937	0,2759	0,3020	0,2816	0,2335	0,1757	0,1209	0,0763	0,0439
	3	0,0105	0,0574	0,1298	0,2013	0,2503	0,2668	0,2522	**0,2150**	0,1665	0,1172
	4	0,0010	0,0112	0,0401	0,0881	0,1460	0,2001	0,2377	0,2508	0,2384	0,2051
	5	0,0001	0,0015	0,0085	0,0264	0,0584	0,1029	0,1536	0,2007	0,2340	0,2461
	6	0,0000	0,0001	0,0012	0,0055	0,0162	0,0368	0,0689	0,1115	0,1596	0,2051
	7	0,0000	0,0000	0,0001	0,0008	0,0031	0,0090	0,0212	0,0425	0,0746	0,1172
	8	0,0000	0,0000	0,0000	0,0001	0,0004	0,0014	0,0043	0,0106	0,0229	0,0439
	9	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001	0,0005	0,0016	0,0042	0,0098
	10	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001	0,0003	0,0010

Valor Esperado e Variância da Distribuição Binomial

Na Seção 5.3, apresentamos fórmulas para calcular o valor esperado e a variância de uma variável aleatória discreta. No caso especial em que a variável aleatória tem uma distribuição binomial com um número conhecido de n ensaios e uma probabilidade conhecida de p sucessos, as fórmulas gerais do valor esperado e variância podem ser simplificadas. Os resultados são apresentados a seguir.

VALOR ESPERADO E VARIÂNCIA DA DISTRIBUIÇÃO BINOMIAL

$$E(x) = \mu = np \qquad \textbf{(5.9)}$$

$$\text{Var}(x) = \sigma^2 = np(1 - p) \qquad \textbf{(5.10)}$$

Figura 5.5 Resultados do Minitab apresentam as probabilidades binomiais relativas ao problema da loja de roupas do Martin

x	P(X = x)
0,00	0,0282
1,00	0,1211
2,00	0,2335
3,00	0,2668
4,00	0,2001
5,00	0,1029
6,00	0,0368
7,00	0,0090
8,00	0,0014
9,00	0,0001
10,00	0,0000

Para o problema com três clientes da loja de roupas do Martin, podemos usar a Equação 5.9 para calcular o número esperado de clientes que farão uma compra.

$$E(x) = np = 3(0,30) = 0,9$$

Suponha que para o próximo mês a loja de roupas do Martin preveja que mil clientes entrarão na loja. Qual é o número esperado de clientes que farão uma compra? A resposta é $\mu = np = (1.000)(0,3) = 300$. Assim, para aumentar o número esperado de vendas, Martin precisa convencer mais clientes a entrar na loja e/ou, de algum modo, aumentar a probabilidade de um cliente individual qualquer fazer uma compra depois de entrar.

Para o problema com três clientes da loja de roupas do Martin, notamos que a variância e o desvio padrão do número de clientes que fazem uma compra são:

$$\sigma^2 = np(1 - p) = 3(0,3)(0,7) = 0,63$$
$$\sigma = \sqrt{0,63} = 0,79$$

Em relação aos mil clientes seguintes que entram na loja, a variância e o desvio padrão do número de clientes que farão uma compra são

$$\sigma^2 = np(1 - p) = 1.000\ (0,3)(0,7) = 210$$
$$\sigma = \sqrt{210} = 14,49$$

NOTAS E COMENTÁRIOS

1. As tabelas binomiais do Apêndice B mostram valores de p somente até $p = 0,50$ inclusive. Poderia parecer que tais tabelas não podem ser usadas quando a probabilidade de sucesso ultrapassa $p = 0,50$. Entretanto, elas podem ser empregadas, notando-se que a probabilidade de $n - x$ fracassos é também a probabilidade de x sucessos. Quando a probabilidade de sucesso é maior que $p = 0,50$, podemos, em substituição, calcular a probabilidade de $n - x$ fracassos. A probabilidade de fracasso, $1 - p$, será menor que 0,50 quando $p > 0,50$.
2. Algumas fontes apresentam tabelas binomiais em forma cumulativa. Ao usarmos tais tabelas precisamos fazer uma subtração para encontrar a probabilidade de x sucessos em n ensaios. Por exemplo, $f(2) = P(x \leq 2) - P(x \leq 1)$. Nossas tabelas fornecem essas probabilidades diretamente. Para calcular probabilidades cumulativas usando nossas tabelas, simplesmente somamos as probabilidades individuais. Por exemplo, para calcular $P(x \leq 2)$ usando nossas tabelas, somamos $f(0) + f(1) + f(2)$.

Exercícios

Métodos

AUTOTESTE

25. Considere um experimento binomial com dois ensaios e $p = 0,4$.
    a. Desenhe um diagrama em árvore desse experimento (ver a Figura 5.3).
    b. Calcule a probabilidade de um sucesso, $f(1)$.
    c. Calcule $f(0)$.
    d. Calcule $f(2)$.
    e. Encontre a probabilidade de pelo menos um sucesso.
    f. Encontre o valor esperado, a variância e o desvio padrão.
26. Considere um experimento binomial com $n = 10$ e $p = 0,10$.
    a. Calcule $f(0)$.
    b. Calcule $f(2)$.
    c. Calcule $P(x \leq 2)$.
    d. Calcule $P(x \geq 1)$.
    e. Calcule $E(x)$.
    f. Calcule $\mathrm{Var}(x)$ e σ.

27. Considere um experimento binomial com $n = 20$ e $p = 0{,}70$.
    a. Calcule $f(12)$.
    b. Calcule $f(16)$.
    c. Calcule $P(x \geq 16)$.
    d. Calcule $P(x \leq 15)$.
    e. Calcule $E(x)$.
    f. Calcule $\text{Var}(x)$ e σ.

Aplicações

28. Uma pesquisa de opinião realizada pela Harris Interactive para à InterContinental Hotels & Resorts perguntou aos entrevistados: "Ao realizar viagens internacionais, você se aventura sozinho para conhecer a cultura local ou se fixa ao seu próprio grupo e itinerários turísticos?" A pesquisa descobriu que 23% dos entrevistados se prendem ao seu grupo turístico (*USA Today*, 21 de janeiro de 2004).
    a. Em uma amostra de seis viajantes internacionais, qual é a probabilidade de dois se prenderem ao seu próprio grupo turístico?
    b. Em uma amostra de seis viajantes internacionais, qual é a probabilidade de pelo menos duas pessoas se prenderem ao seu próprio grupo turístico?
    c. Em uma amostra de dez viajantes internacionais, qual é a probabilidade de nenhum se prender ao seu próprio grupo turístico?
29. De acordo com uma pesquisa de opinião realizada pela *Business Week*/Harris Poll entre 1.035 adultos, 40% dos entrevistados concordaram fortemente com a proposição de que os negócios têm muita influência sobre o estilo de vida dos norte-americanos (*Business Week*, 11 de setembro de 2000). Considere essa porcentagem como representativa da população norte-americana. Em uma amostra de 20 indivíduos, tomada em determinado instante da população norte-americana, qual é a probabilidade de pelo menos cinco indivíduos acharem que os negócios têm muito mais influência sobre o estilo de vida norte-americano?

AUTOTESTE

30. Quando uma máquina nova funciona adequadamente, somente 3% dos itens produzidos apresentam defeitos. Suponha escolhermos aleatoriamente duas peças produzidas na máquina e estarmos interessados no número de peças defeituosas encontradas.
    a. Descreva as condições sob as quais essa situação seria um experimento binomial.
    b. Desenhe um diagrama em árvore similar à Figura 5.3, ilustrando esse problema como um experimento de dois ensaios.
    c. Quantos resultados experimentais resultam em encontrarmos exatamente um defeito?
    d. Calcule as probabilidades de não encontrarmos defeitos, encontrarmos exatamente um defeito e encontrarmos dois defeitos.
31. Nove por cento dos estudantes universitários portam cartões de crédito com limites maiores que US$ 7 mil (*Reader's Digest*, julho de 2002). Suponha que dez estudantes universitários sejam escolhidos aleatoriamente para serem entrevistados acerca do uso do cartão de crédito.
    a. A escolha dos dez estudantes é um experimento binomial? Explique.
    b. Qual é a probabilidade de dois dos estudantes terem um limite de crédito maior que US$ 7 mil?
    c. Qual é a probabilidade de nenhum ter limite de crédito maior que US$ 7 mil?
    d. Qual é a probabilidade de pelo menos três terem limites de crédito maiores que US$ 7 mil?
32. Os sistemas militares de radar e de mísseis são concebidos para um país precaver-se de ataques inimigos. Uma questão de confiabilidade é se um sistema de detecção será capaz de identificar um ataque e disparar um alarme. Considere que determinado sistema de detecção tenha uma probabilidade de 0,90 de detectar um ataque de mísseis. Use a distribuição de probabilidade binomial para responder às seguintes questões:
    a. Qual é a probabilidade de um único sistema de detecção detectar um ataque?
    b. Se dois sistemas de detecção estão instalados na mesma área e operam independentemente, qual é a probabilidade de pelo menos um dos sistemas detectar o ataque?
    c. Se três sistemas estão instalados, qual é a probabilidade de que pelo menos um dos sistemas detectar o ataque?
    d. Você recomendaria o uso de múltiplos sistemas de detecção? Explique.

33. Cinqüenta por cento dos norte-americanos acreditavam que o país se encontrava em recessão, não obstante, tecnicamente, a economia não apresentar dois semestres seguidos de crescimento negativo (*Business Week*, 30 de julho de 2001). Em relação a uma amostra de 20 norte-americanos, faça os seguintes cálculos.
    a. Calcule a probabilidade de exatamente 12 pessoas acreditarem que o país se encontrava em recessão.
    b. Calcule a probabilidade de não mais que cinco pessoas acreditarem que o país se encontrava em recessão.
    c. Quantos pessoas você acha que diriam que o país se encontrava em recessão?
    d. Calcule a variância e o desvio padrão do número de pessoas que acreditavam que o país se encontrava em recessão?
34. Quarenta por cento das pessoas que viajam a negócios portam um telefone celular ou um *laptop* (*USA Today*, 12 de setembro de 2000). Em relação a uma amostra de 15 pessoas que viajam a negócios, faça os seguintes cálculos.
    a. Calcule a probabilidade de três dos viajantes portarem um telefone celular ou um *laptop*.
    b. Calcule a probabilidade de 12 dos viajantes não portarem telefone celular nem *laptop*.
    c. Calcule a probabilidade de pelo menos três dos viajantes portarem um telefone celular ou um *laptop*.
35. Uma universidade descobriu que 20% dos seus estudantes saem sem concluir o curso introdutório de estatística. Considere que 20 estudantes tenham se matriculado para o curso.
    a. Calcule a probabilidade de dois ou menos desistirem.
    b. Calcule a probabilidade de exatamente quatro desistirem.
    c. Calcule a probabilidade de mais de três desistirem.
    d. Calcule o número esperado de desistências.
36. Para o caso especial de uma variável aleatória binomial, estabelecemos que a variância poderia ser calculada por meio da fórmula $\sigma^2 = np(1-p)$. Em relação ao problema da loja de roupas do Martin, considerando $n = 3$ e $p = 0{,}3$, encontramos $\sigma^2 = np(1-p) = 3(0{,}3)(0{,}7) = 0{,}63$. Use a definição geral de variância de uma variável aleatória discreta dada pela Equação 5.5 e as probabilidades apresentadas na Tabela 5.7 para verificar se a variância é realmente 0,63.
37. Setenta e dois por cento dos norte-americanos têm acesso on-line (*CNBC*, 3 de dezembro de 2001). Em uma amostra aleatória de 30 pessoas, qual é o número esperado de pessoas com acesso on-line? Qual é a variância e o desvio padrão?

5.5 DISTRIBUIÇÃO DE POISSON

Nesta seção, consideraremos uma variável aleatória discreta que muitas vezes é útil para calcular o número de ocorrências ao longo de um intervalo de tempo ou espaço específicos. Por exemplo, a variável aleatória de interesse pode ser o número de carros que chegam a um lava-rápido em uma hora, o número de reparos necessários em 16 quilômetros de uma rodovia ou o número de vazamentos em 160 quilômetros de tubulação. Se as duas propriedades seguintes forem satisfeitas, o número de ocorrências será uma variável aleatória descrita pela **função probabilidade de Poisson**.

A distribuição de probabilidade de Poisson freqüentemente é usada para traçar um modelo de chegadas aleatórias em situações que recorrem a filas de espera.

PROPRIEDADES DE UM EXPERIMENTO DE POISSON

1. A probabilidade de uma ocorrência é a mesma para dois intervalos quaisquer de igual comprimento.
2. A ocorrência ou não-ocorrência em determinado intervalo é independente da ocorrência ou não-ocorrência em outro intervalo.

Siméon Poisson lecionou matemática na Ecole Polytechnique de Paris, de 1802 a 1808. Em 1837, ele publicou uma obra intitulada *Pesquisa Sobre a Probabilidade de Veredictos Civis e Criminais*, a qual inclui uma discussão daquilo que mais tarde passaria a ser conhecido como distribuição de Poisson.

A **função probabilidade de Poisson** é definida pela Equação (5.11).

FUNÇÃO DE PROBABILIDADE DE POISSON

$$f(x) = \frac{\mu^x e^{-\mu}}{x!} \qquad \textbf{(5.11)}$$

em que

$f(x)$ = a probabilidade de x ocorrências em um intervalo
μ = valor esperado, ou número médio, de ocorrências
e = 2,71828

Antes de considerarmos um exemplo específico para verificar como a distribuição de Poisson pode ser aplicada, observe que o número de ocorrências, x, não tem limites máximos. Ela é uma variável aleatória discreta que pode assumir uma seqüência infinita de valores ($x = 0, 1, 2, ...$).

Um Exemplo Envolvendo Intervalos de Tempo

A Bell Labs utiliza a distribuição de Poisson para traçar um modelo da chegada de chamadas telefônicas.

Suponha que estejamos interessados no número de carros que chegam a um caixa automático *drive-thru* de um banco durante um período de 15 minutos nas manhãs de fins de semana. Se considerarmos que a probabilidade de um carro chegar é a mesma para dois períodos quaisquer de igual duração e que o fato de carros chegarem ou não chegarem em qualquer período é independente da chegada ou não-chegada de outro em qualquer outro período, a função probabilidade de Poisson é aplicável. Considere que essas hipóteses sejam satisfeitas e que a análise dos dados históricos mostre que o número médio de carros que chegam no período de 15 minutos é 10; sendo assim, aplica-se a seguinte função probabilidade:

$$f(x) = \frac{10^x e^{-10}}{x!}$$

A variável aleatória nesse caso é x = o número de carros que chegam em um período de 15 minutos qualquer.

Se a gerência quisesse saber a probabilidade de exatamente cinco carros chegarem em 15 minutos, definiríamos $x = 5$ e, desse modo, obteríamos

$$\text{Probabilidade de exatamente 5 carros chegarem em 15 minutos} = f(5) = \frac{10^5 e^{-10}}{5!} = 0{,}0378$$

Embora essa probabilidade tenha sido determinada calculando-se a função probabilidade com $\mu = 10$ e $x = 5$, muitas vezes é mais fácil consultar uma tabela para verificar a distribuição de Poisson. Uma tabela fornece probabilidades para valores específicos de x e de μ. Incluímos esse tipo de tabela no Apêndice B com o título de Tabela 7. Por conveniência, reproduzimos parte dessa tabela com o título de Tabela 5.9. Observe que para usarmos a tabela de probabilidades de Poisson precisamos conhecer somente os valores de x e de μ. Da Tabela 5.9, sabemos que a probabilidade de chegarem cinco carros em um período de 15 minutos é calculada encontrando-se o valor na linha da tabela correspondente a $x = 5$ e a coluna da tabela correspondente a $\mu = 10$. Portanto, obtemos $f(5) = 0{,}0378$.

Uma propriedade da distribuição de Poisson é que a média e a variância são iguais.

No exemplo anterior, a média da distribuição de Poisson é $\mu = 10$ carros que chegam por período de 15 minutos. Uma propriedade da distribuição de Poisson é que a média da distribuição e a variância da distribuição são *iguais*. Sendo assim, a variância do número de carros que chegam durante períodos de 15 minutos é $\sigma = 10$. O desvio padrão é $\sigma = \sqrt{10} = 3{,}16$.

Nossa ilustração envolve um período de 15 minutos, mas outros períodos podem ser usados. Suponha que queiramos computar a probabilidade de um carro chegar em um período de três minutos. Uma vez que 10 é o número esperado de carros que chegam em um período de 15 minutos, observamos que 10/15 = 2/3 é o número esperado de carros que chegam em um período de um minuto e que (2/3)(3 minutos) = 2 é o número esperado de carros que chegam em um período de três minutos. Assim, a probabilidade de x carros chegarem em um período de três minutos, com $\mu = 2$, é dada pela seguinte função probabilidade de Poisson:

$$f(x) = \frac{2^x e^{-2}}{x!}$$

Tabela 5.9 Valores selecionados de tabelas de probabilidade de Poisson
Exemplo: $\mu = 10$, $x = 5$; $f(5) = 0,0378$

					μ					
x	9,1	9,2	9,3	9,4	9,5	9,6	9,7	9,8	9,9	10
0	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0000
1	0,0010	0,0009	0,0009	0,0008	0,0007	0,0007	0,0006	0,0005	0,0005	0,0005
2	0,0046	0,0043	0,0040	0,0037	0,0034	0,0031	0,0029	0,0027	0,0025	0,0023
3	0,0140	0,0131	0,0123	0,0115	0,0107	0,0100	0,0093	0,0087	0,0081	0,0076
4	0,0319	0,0302	0,0285	0,0269	0,0254	0,0240	0,0226	0,0213	0,0201	0,0189
5	0,0581	0,0555	0,0530	0,0506	0,0483	0,0460	0,0439	0,0418	0,0398	0,0378
6	0,0881	0,0851	0,0822	0,0793	0,0764	0,0736	0,0709	0,0682	0,0656	0,0631
7	0,1145	0,1118	0,1091	0,1064	0,1037	0,1010	0,0982	0,0955	0,0928	0,0901
8	0,1302	0,1286	0,1269	0,1251	0,1232	0,1212	0,1191	0,1170	0,1148	0,1126
9	0,1317	0,1315	0,1311	0,1306	0,1300	0,1293	0,1284	0,1274	0,1263	0,1251
10	0,1198	0,1210	0,1219	0,1228	0,1235	0,1241	0,1245	0,1249	0,1250	0,1251
11	0,0991	0,1012	0,1031	0,1049	0,1067	0,1083	0,1098	0,1112	0,1125	0,1137
12	0,0752	0,0776	0,0799	0,0822	0,0844	0,0866	0,0888	0,0908	0,0928	0,0948
13	0,0526	0,0549	0,0572	0,0594	0,0617	0,0640	0,0662	0,0685	0,0707	0,0729
14	0,0342	0,0361	0,0380	0,0399	0,0419	0,0439	0,0459	0,0479	0,0500	0,0521
15	0,0208	0,0221	0,0235	0,0250	0,0265	0,0281	0,0297	0,0313	0,0330	0,0347
16	0,0118	0,0127	0,0137	0,0147	0,0157	0,0168	0,0180	0,0192	0,0204	0,0217
17	0,0063	0,0069	0,0075	0,0081	0,0088	0,0095	0,0103	0,0111	0,0119	0,0128
18	0,0032	0,0035	0,0039	0,0042	0,0046	0,0051	0,0055	0,0060	0,0065	0,0071
19	0,0015	0,0017	0,0019	0,0021	0,0023	0,0026	0,0028	0,0031	0,0034	0,0037
20	0,0007	0,0008	0,0009	0,0010	0,0011	0,0012	0,0014	0,0015	0,0017	0,0019
21	0,0003	0,0003	0,0004	0,0004	0,0005	0,0006	0,0006	0,0007	0,0008	0,0009
22	0,0001	0,0001	0,0002	0,0002	0,0002	0,0002	0,0003	0,0003	0,0004	0,0004
23	0,0000	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0002	0,0002
24	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001	0,0001	0,0001

A probabilidade de um carro chegar em um período de três minutos é calculada da seguinte maneira:

$$\begin{array}{l}\text{Probabilidade de exatamente}\\ \text{1 carro chegar em 3 minutos}\end{array} = f(1) = \frac{2^1 e^{-2}}{1!} = 0{,}2707$$

Calculamos anteriormente a probabilidade de cinco carros chegarem em um período de 15 minutos; foi 0,0378. Observe que a probabilidade de um carro chegar em um período de três minutos (0,2707) não é a mesma. Quando se calcula uma probabilidade de Poisson para um intervalo de tempo diferente, devemos primeiramente converter a taxa média de chegada para o período de interesse e depois calcular a probabilidade.

Um Exemplo Envolvendo Intervalos de Comprimento ou de Distância

Vamos ilustrar uma aplicação que não envolve intervalos de tempo na qual a distribuição de probabilidade de Poisson é útil. Suponha estarmos preocupados com a ocorrência de defeitos importantes em uma rodovia um mês depois do recapeamento. Vamos supor que a probabilidade de um defeito seja a mesma em dois intervalos quaisquer de igual extensão na rodovia e que a ocorrência ou não-ocorrência de um defeito em determinado intervalo seja independente da ocorrência ou não-ocorrência de um defeito em outro intervalo qualquer. Assim, a distribuição de probabilidade de Poisson pode ser aplicada.

Suponha que saibamos que defeitos importantes ocorrem um mês depois do recapeamento à taxa média de dois defeitos por quilômetro. Vamos encontrar a probabilidade de não haver nenhum defeito importante em um trecho de três quilômetros, em especial, na rodovia. Como estamos interessados em um intervalo com uma extensão de três quilômetros, μ = (2 defeitos/quilômetro)(3 quilômetros) = 6 representa o número esperado de defeitos importantes no trecho de três quilômetros da rodovia. Usando a Equação 5.11, observamos que a probabilidade de não-ocorrência de defeitos importantes é $f(0) = 6^0 e^{-6}/0! = 0{,}0025$. Assim, é improvável que nenhum defeito importante ocorra no trecho de três quilômetros. Realmente, esse exemplo indica uma probabilidade de 1 – 0,0025 = 0,9975 de pelo menos um defeito importante ocorrer em um trecho da rodovia.

Exercícios

Métodos

38. Considere uma distribuição de Poisson com $\mu = 3$.
    a. Escreva a função probabilidade de Poisson apropriada.
    b. Encontre $f(2)$.
    c. Encontre $f(1)$.
    d. Encontre $P(x \geq 2)$.

39. Considere uma distribuição de Poisson com um número médio de duas ocorrências por período.
    a. Escreva a função probabilidade de Poisson apropriada.
    b. Qual é o número esperado de ocorrências em três períodos?
    c. Escreva a função probabilidade de Poisson apropriada para determinar a probabilidade de x ocorrências em três períodos.
    d. Encontre a probabilidade de duas ocorrências em um período.
    e. Encontre a probabilidade de seis ocorrências em três períodos.
    f. Encontre a probabilidade de cinco ocorrências em dois períodos.

AUTOTESTE

Aplicações

40. Chamadas telefônicas são recebidas à taxa de 48 por hora no balcão de reservas da Regional Airways.
    a. Calcule a probabilidade de receberem três chamadas em um intervalo de tempo de cinco minutos.
    b. Calcule a probabilidade de receberem exatamente dez chamadas em 15 minutos.
    c. Suponha não haver nenhuma chamada em espera no momento. Se o recepcionista demora cinco minutos para completar a chamada atual, quantas ligações você acha que permanecerão em espera nesse tempo? Qual é a probabilidade de não haver nenhuma ligação em espera?
    d. Se nenhuma chamada está em processamento neste momento, qual é a probabilidade de o recepcionista ter três minutos de tempo pessoal sem ser interrompido?

41. Durante o período em que uma universidade local recebe inscrições por telefone, as chamadas telefônicas são recebidas a uma taxa de uma ligação a cada dois minutos.
    a. Qual é o número esperado de ligações recebidas em uma hora?
    b. Qual é a probabilidade de três ligações serem recebidas em cinco minutos?
    c. Qual é a probabilidade de nenhuma ligação ser recebida em um período de cinco minutos?

42. Os estabelecimentos da Bed & Breakfast (B&B) registraram a estada de mais de 50 milhões de hóspedes no ano passado. O site da Bed and Breakfast Inns of North America (www.bestinns.net), o qual tem uma média de aproximadamente sete visitas por minuto, possibilita a muitos estabelecimentos da B&B atraírem hóspedes sem a necessidade de esperar vários anos para serem citados em guias de viagem (*Time*, setembro de 2001).
    a. Calcule a probabilidade de não haver nenhuma visita ao site no período de um minuto.
    b. Calcule a probabilidade de haver duas ou mais visitas ao site no período de um minuto.
    c. Calcule a probabilidade de haver uma ou mais visitas ao site em um período de 30 segundos.
    d. Calcule a probabilidade de haver cinco ou mais visitas ao site no período de um minuto.

AUTOTESTE

43. Os passageiros de uma empresa aérea chegam aleatória e independentemente ao balcão de controle de passageiros de um importante aeroporto internacional. A taxa média de chegada são 10 passageiros por minuto.
    a. Calcule a probabilidade de ninguém chegar no período de um minuto.
    b. Calcule a probabilidade de três ou menos passageiros chegarem no período de um minuto.
    c. Calcule a probabilidade de ninguém chegar em um período de 15 segundos.
    d. Calcule a probabilidade de pelo menos um passageiro chegar em um período de 15 segundos.

44. De 1990 a 1999 houve uma média de aproximadamente 26 acidentes aeronáuticos por ano que acarretaram a morte de um ou mais passageiros. A partir de 2000, a média decresceu para 15 acidentes por ano (*The World Almanac and Book of Facts*, 2004). Suponha que os acidentes aeronáuticos continuem a ocorrer à taxa de 15 acidentes por ano.

a. Calcule o número médio de acidentes aeronáuticos por mês.
b. Calcule a probabilidade de não ocorrer nenhum acidente durante um mês.
c. Calcule a probabilidade de ocorrer exatamente um acidente durante um mês.
d. Calcule a probabilidade de ocorrer mais de um acidente durante um mês.

45. O National Safety Council registrou que as mortes relacionadas ao uso de *air-bags* caíram para 18 no ano 2000 (http://www.nsc.org).
    a. Calcule o número esperado de mortes relacionadas ao uso de *air-bags* por mês.
    b. Calcule a probabilidade de não ocorrer nenhuma morte relacionada ao uso de *air-bags* em um mês.
    c. Calcule a probabilidade de ocorrer duas ou mais mortes relacionadas ao uso de *air-bags* em um mês.

5.6 DISTRIBUIÇÃO DE PROBABILIDADE HIPERGEOMÉTRICA

A **distribuição de probabilidade hipergeométrica** relaciona-se restritamente com a distribuição de probabilidade binomial. As duas distribuições de probabilidade diferem sob dois aspectos fundamentais. Quando se trata da distribuição hipergeométrica, os ensaios não são independentes e a probabilidade de sucesso se modifica de ensaio a ensaio.

Na notação usual da distribuição de probabilidade hipergeométrica, r denota o número de elementos da população de tamanho N que são rotulados de sucesso e $N - r$ denota o número de elementos da população que são rotulados de fracasso. A **função probabilidade hipergeométrica** é usada para calcular a probabilidade de obtermos x elementos rotulados de sucesso e $n - x$ elementos rotulados de fracasso em uma seleção aleatória de n elementos, selecionados sem substituição. Para que isso ocorra, precisamos obter x sucessos dos r sucessos na população e $n - x$ fracassos dos $N - r$ fracassos. A seguinte função probabilidade hipergeométrica fornece $f(x)$, a qual é a probabilidade de obtermos x sucessos em uma amostra de tamanho n.

FUNÇÃO PROBABILIDADE HIPERGEOMÉTRICA

$$f(x) = \frac{\binom{r}{x}\binom{N-r}{n-x}}{\binom{N}{n}} \quad \text{para } 0 \leq x \leq r \qquad (5.12)$$

em que

$f(x)$ = probabilidade de x sucessos em n ensaios
n = número de ensaios
N = número de elementos da população
r = número de elementos da população rotulados de sucesso

Observe que $\binom{N}{n}$ representa o número de maneiras pelas quais uma amostra de tamanho n pode ser selecionada de uma população de tamanho N; $\binom{r}{x}$ representa o número de maneiras pelas quais x sucessos podem ser selecionados de um total de r sucessos na população; e $\binom{N-r}{n-x}$ representa o número de maneiras pelas quais $n - x$ fracassos pode ser selecionado de um total de $N - r$ fracassos na população.

Para ilustrar os cálculos envolvidos no uso da Equação 5.12, consideremos a seguinte aplicação de controle da qualidade. Os fusíveis elétricos produzidos pela Ontario Electric são embalados em caixas de 12 unidades cada uma. Suponha que um controlador da qualidade selecione aleatoriamente três dos 12 fusíveis contidos em uma caixa para testá-los. Se a caixa contém exatamente cinco fusíveis defeituosos, qual é a probabilidade de o controlador da qualidade encontrar exatamente um dos três fusíveis defeituosos? Nessa aplicação, $n = 3$ e $N = 12$. Com $r = 5$ fusíveis defeituosos na caixa, a probabilidade de encontrar $x = 1$ fusível defeituoso é:

$$f(1) = \frac{\binom{5}{1}\binom{7}{2}}{\binom{12}{3}} = \frac{\left(\frac{5!}{1!4!}\right)\left(\frac{7!}{2!5!}\right)}{\left(\frac{12!}{3!9!}\right)} = \frac{(5)(21)}{220} = 0{,}4773$$

Suponha agora que queiramos saber qual é a probabilidade de encontrar *pelo menos* um fusível defeituoso. A maneira mais fácil de responder a essa questão é calcular primeiramente a probabilidade de o controlador da qualidade não encontrar nenhum fusível defeituoso. A probabilidade de $x = 0$ é

$$f(0) = \frac{\binom{5}{0}\binom{7}{3}}{\binom{12}{3}} = \frac{\left(\frac{5!}{0!5!}\right)\left(\frac{7!}{3!4!}\right)}{\left(\frac{12!}{3!9!}\right)} = \frac{(1)(35)}{220} = 0{,}1591$$

Com a probabilidade de não haver nenhum fusível defeituoso $f(0) = 0{,}1591$, concluímos que a probabilidade de encontrar pelo menos um fusível defeituoso deve ser $1 - 0{,}1591 = 0{,}8409$. Assim, há a probabilidade razoavelmente elevada de o controlador da qualidade vir a encontrar pelo menos um fusível defeituoso.

A média e a variância de uma distribuição hipergeométrica são apresentadas a seguir.

$$E(x) = \mu = n\left(\frac{r}{N}\right) \qquad \textbf{(5.13)}$$

$$\text{Var}(x) = \sigma^2 = n\left(\frac{r}{N}\right)\left(1 - \frac{r}{N}\right)\left(\frac{N-n}{N-1}\right) \qquad \textbf{(5.14)}$$

No exemplo anterior, $n = 3$, $r = 5$ e $N = 12$. Assim, a média e a variância do número de fusíveis defeituosos é

$$\mu = n\left(\frac{r}{N}\right) = 3\left(\frac{5}{12}\right) = 1{,}25$$

$$\sigma^2 = n\left(\frac{r}{N}\right)\left(1 - \frac{r}{N}\right)\left(\frac{N-n}{N-1}\right) = 3\left(\frac{5}{12}\right)\left(1 - \frac{5}{12}\right)\left(\frac{12-3}{12-1}\right) = 0{,}60$$

O desvio padrão é $\sigma = \sqrt{0{,}60} = 0{,}77$.

NOTAS E COMENTÁRIOS

Considere uma distribuição hipergeométrica com n ensaios. Digamos que $p = (r/N)$ denote a probabilidade de um sucesso no primeiro ensaio. Se o tamanho da população for grande, o termo $(N - n)/(N - 1)$ da Equação 5.14 aproxima-se de 1. Em conseqüência, o valor esperado e a variância podem ser escritos como $E(x) = np$ e $\text{Var}(x) = np(1 - p)$. Note que essas expressões são similares àquelas usadas para calcular o valor esperado e a variância de uma distribuição binomial, como nas Equações 5.9 e (5.10). Quando o tamanho da população é grande, uma distribuição hipergeométrica pode ser aproximada por meio de uma distribuição binomial com n ensaios e uma probabilidade de $p = (r/N)$.

Exercícios

Métodos

46. Suponha $N = 10$ e $r = 3$. Calcule as probabilidades hipergeométricas para os seguintes valores de n e x.
    a. $n = 4, x = 1$.
    b. $n = 2, x = 2$.
    c. $n = 2, x = 0$.
    d. $n = 4, x = 2$.
47. Suponha $N = 15$ e $r = 4$. Qual é a probabilidade de $x = 3$ para $n = 10$?

Aplicações

48. Em uma pesquisa de opinião realizada pela Gallup Organization foi feita a seguinte pergunta aos entrevistados: "A qual esporte você prefere assistir?" O futebol e o basquete classificaram-se em primeiro e segundo lugares, respectivamente, em termos de preferência (http://www.gallup.com, 3 de janeiro de 2004). Suponha que em um grupo de dez pessoas, sete prefiram futebol e três, basquete. Uma amostra aleatória de três dessas pessoas é selecionada.

AUTOTESTE

    a. Qual é a probabilidade de exatamente duas preferirem futebol?
    b. Qual é a probabilidade de a maioria (duas ou três) preferir futebol?

49. O *blackjack*, ou vinte-e-um como é freqüentemente chamado, é um jogo de azar popular jogado nos cassinos de Las Vegas. O jogador recebe duas cartas. As cartas da corte (valete, dama e rei) e os dez valem dez pontos. Os ases valem um ou 11 pontos. Um baralho de 52 cartas contém 16 cartas que valem dez pontos (valetes, reis, damas e dez) e quatro ases.
    a. Qual é a probabilidade de ambas as cartas tiradas serem ases ou cartas de dez pontos?
    b. Qual é a probabilidade de ambas as cartas serem ases?
    c. Qual é a probabilidade de ambas as cartas valerem dez pontos?
    d. Um *blackjack* forma-se com uma carta de dez pontos e um ás, totalizando 21 pontos. Use suas respostas às questões (a), (b) e (c) para determinar a probabilidade de um jogador tirar um *blackjack*. (*Dica*: A questão (d) não é um problema hipergeométrico. Desenvolva sua própria relação lógica de como as probabilidades hipergeométricas dos itens (a), (b) e (c) podem ser combinadas para responder a essa questão).

50. A Axline Computers produz computadores pessoais em duas fábricas: uma no Texas e outra no Havaí. A fábrica do Texas tem 40 empregados e a do Havaí, 20. Pede-se a uma amostra aleatória de dez empregados para preencherem um questionário de benefícios.
    a. Qual é a probabilidade de nenhum dos empregados da amostra trabalhar na fábrica do Havaí?
    b. Qual é a probabilidade de um dos empregados da amostra trabalhar na fábrica do Havaí?
    c. Qual é a probabilidade de dois empregados ou mais dos empregados da amostra trabalharem na fábrica do Havaí?
    d. Qual é a probabilidade de nove dos empregados da amostra trabalharem na fábrica do Texas?

51. A *2003 Zagat Restaurant Survey* fornece classificações referentes à qualidade da comida, conforto e atendimento de alguns dos grandes restaurantes nos Estados Unidos. Para os 15 restaurantes mais bem classificados localizados em Boston o preço médio de um jantar, incluindo uma bebida e a gorjeta, era US$ 48,60. Você chega a Boston em uma viagem de negócios e jantará em três desses restaurantes. Sua empresa lhe reembolsará um valor máximo de US$ 50 por jantar. Seus colegas de negócios que têm familiaridade com esses restaurantes disseram-lhe que o custo das refeições em um terço dos restaurantes ultrapassará o valor de US$ 50. Suponha que você escolha aleatoriamente três desses restaurantes para fazer suas refeições.

AUTOTESTE

    a. Qual é a probabilidade de nenhuma das refeições ultrapassar o custo coberto por sua empresa?
    b. Qual é a probabilidade de uma das refeições ultrapassar o custo coberto por sua empresa?
    c. Qual é a probabilidade de duas ou mais refeições ultrapassarem o custo coberto por sua empresa?
    d. Qual é a probabilidade de todas as três refeições ultrapassarem o custo coberto por sua empresa?

52. Uma remessa de dez itens contém duas unidades com defeito e oito unidades sem defeito. Na inspeção de embarque, uma amostra de unidades será selecionada e testada. Se uma unidade com defeito for encontrada, a remessa de dez unidades será rejeitada.
    a. Se uma amostra de três itens for selecionada, qual é a probabilidade de o embarque ser rejeitado?
    b. Se uma amostra de quatro itens for selecionada, qual é a probabilidade de o embarque ser rejeitado?
    c. Se uma amostra de cinco itens for selecionada, qual é a probabilidade de o embarque ser rejeitado?
    d. Se a administração quiser obter uma probabilidade de 0,90 de rejeição de um embarque com duas unidades defeituosas e oito unidades sem defeito, qual seria o tamanho da amostra por você recomendada?

Resumo

Uma variável aleatória constitui uma descrição numérica do resultado de um experimento. A distribuição de probabilidade de uma variável aleatória descreve a maneira pela qual as probabilidades se distribuem ao longo dos valores que a variável aleatória pode assumir. Para qualquer variável aleatória discreta x a distribuição de probabilidade é definida por uma função probabilidade, denotada por $f(x)$ que fornece a probabilidade associada a cada valor da variável aleatória. Uma vez que a função probabilidade tenha sido definida, podemos calcular o valor esperado, a variância e o desvio padrão da variável aleatória.

A distribuição de probabilidade binomial pode ser usada para determinar a probabilidade de x sucessos em n ensaios sempre que o experimento apresentar as seguintes propriedades:

1. O experimento consiste em uma seqüência de n ensaios idênticos.
2. Dois resultados são possíveis em cada um dos ensaios, sendo um deles chamado sucesso e o outro, fracasso.
3. A probabilidade de um sucesso p não se modifica de ensaio a ensaio. Conseqüentemente, a probabilidade de fracasso, $1 - p$, não se modifica de ensaio a ensaio.
4. Os ensaios são independentes.

Quando as quatro condições são válidas a função probabilidade binomial pode ser usada para determinar a probabilidade de se obter x sucessos em n ensaios. Também foram apresentadas fórmulas relativas à média e à variância da distribuição binomial.

A distribuição de Poisson é usada quando é desejável determinar a probabilidade de se obter x ocorrências ao longo de um intervalo de tempo ou de espaço. As seguintes hipóteses são necessárias para que a distribuição de Poisson seja aplicável.

1. A probabilidade de uma ocorrência do evento é a mesma para dois intervalos quaisquer de igual comprimento.
2. A ocorrência ou não-ocorrência do evento em qualquer intervalo é independente da ocorrência ou não-ocorrência do evento em qualquer outro intervalo.

Uma terceira distribuição de probabilidade discreta, a hipergeométrica, foi apresentada na Seção 5.6. À semelhança da distribuição binomial, ela é usada para calcular a probabilidade de x sucessos em n ensaios. Mas, em comparação com a binomial, a probabilidade de sucesso modifica-se de ensaio a ensaio.

Glossário

Variável aleatória Uma descrição numérica do resultado de um experimento.

Variável aleatória discreta Uma variável aleatória que pode assumir ou um número finito de valores ou uma seqüência de valores infinitos.

Variável aleatória contínua Uma variável aleatória que pode assumir qualquer valor numérico em um intervalo ou grupo de intervalos.

Distribuição de probabilidade Uma descrição de como as probabilidades se distribuem ao longo dos valores da variável aleatória.

Função probabilidade Uma função, denotada por $f(x)$, que fornece a probabilidade de x assumir um valor determinado para uma variável aleatória discreta.

Distribuição de probabilidade discreta uniforme Uma distribuição de probabilidade para a qual cada valor possível da variável aleatória tem a mesma probabilidade.

Valor esperado Uma medida da posição central de uma variável aleatória.

Variância Uma medida da variabilidade, ou dispersão, de uma variável aleatória.

Desvio padrão A raiz quadrada positiva da variância.

Experimento binomial Um experimento que possui as quatro propriedades definidas no início da Seção 5.4.

Distribuição de probabilidade binomial A distribuição de probabilidade que mostra a probabilidade de x sucessos em n ensaios de um experimento binomial.

Função probabilidade binomial A função usada para calcular probabilidades binomiais.

Distribuição de probabilidade de Poisson Uma distribuição de probabilidade que mostra a probabilidade de x ocorrências de um evento em um intervalo de tempo ou de espaço específicos.

Função probabilidade de Poisson A função usada para calcular as probabilidades de Poisson.

Distribuição de probabilidade hipergeométrica Uma distribuição de probabilidade que mostra a probabilidade de x sucessos em n ensaios de uma população com r sucessos e $N - r$ fracassos.

Função probabilidade hipergeométrica A função usada para calcular probabilidades hipergeométricas.

Fórmulas-Chave

Função Discreta Uniforme de Probabilidade

$$f(x) = 1/n \tag{5.3}$$

em que

n = o número de valores que a variável aleatória pode assumir

Valor Esperado de uma Variável Aleatória Discreta

$$E(x) = \mu = \Sigma x f(x) \tag{5.4}$$

Variância de uma Variável Aleatória Discreta

$$\text{Var}(x) = \sigma^2 = \Sigma(x - \mu)^2 f(x) \tag{5.5}$$

Número de Resultados Experimentais que Fornecem Exatamente x Sucessos em n Ensaios

$$\binom{n}{x} = \frac{n!}{x!(n - x)!} \tag{5.6}$$

Função Probabilidade Binomial

$$f(x) = \binom{n}{x} p^x (1 - p)^{(n-x)} \tag{5.8}$$

Valor Esperado da Distribuição Binomial

$$E(x) = \mu = np \tag{5.9}$$

Variância da Distribuição Binomial

$$\text{Var}(x) = \sigma^2 = np(1 - p) \tag{5.10}$$

Função Probabilidade de Poisson

$$f(x) = \frac{\mu^x e^{-\mu}}{x!} \tag{5.11}$$

Função Probabilidade Hipergeométrica

$$f(x) = \frac{\binom{r}{x}\binom{N - r}{n - x}}{\binom{N}{n}} \quad \text{para } 0 \leq x \leq r \tag{5.12}$$

Valor Esperado da Distribuição Hipergeométrica

$$E(x) = \mu = n\left(\frac{r}{N}\right) \tag{5.13}$$

Variância da Distribuição Hipergeométrica

$$\text{Var}(x) = \sigma^2 = n\left(\frac{r}{N}\right)\left(1 - \frac{r}{N}\right)\left(\frac{N - n}{N - 1}\right) \tag{5.14}$$

Exercícios Suplementares

53. A Big Money Poll realizada pela *Barron's* perguntou a 131 gerentes de investimento de várias partes dos Estados Unidos a respeito de suas previsões de investimento no curto prazo (*Barron's*, 28 de outubro de 2002). Suas respostas mostraram que 4% eram muito altistas (*bullish*), 39% eram altistas, 29% eram neutros, 21% eram baixistas (*bearish*) e 7% eram muito baixistas. Seja x a variável aleatória que reflete o nível de otimismo em relação ao mercado. Estabeleça $x = 5$ para os muito altistas e $x = 1$ para os muito baixistas.
    a. Desenvolva a distribuição de probabilidade correspondente ao nível de otimismo dos gerentes de investimento.
    b. Calcule o valor esperado do nível de otimismo.
    c. Calcule a variância e o desvio padrão do nível de otimismo.
    d. Comente o que implicam os seus resultados em relação ao nível de otimismo e sua variabilidade.
54. A American Association of Individual Investors publica um guia anual dos principais fundos mútuos de investimentos (*The Individual Investor's Guide to the Top Mutual Funds*, 22e, American Association of Individual Investors, 2003). A Tabela 5.10 contém suas classificações do risco total, referentes a 29 categorias de fundos mútuos de investimentos.
    a. Admita $x = 1$ para baixo risco e $x = 5$ para alto risco e desenvolva uma distribuição de probabilidade para o nível de risco.
    b. Quais são o valor esperado e a variância do risco total?
    c. Ocorre que 11 das categorias de fundos eram fundos de debêntures. Dos fundos de debêntures, sete categorias tinham uma baixa classificação e quatro tinham uma classificação abaixo da média. Compare o risco total dos fundos de debêntures com as 18 categorias de fundos de ações.

Tabela 5.10 Classificação do risco para 29 categorias de fundos mútuos de investimentos

Risco total	Número de Categorias de Fundos
Baixo	7
Abaixo da média	6
Médio	3
Acima da média	6
Elevado	7

55. O processo de elaboração orçamentária de uma universidade do meio-oeste resultou em previsões de gastos para o ano vindouro equivalentes a (em milhões) US$ 9, US$ 10, US$ 11, US$ 12 e US$ 13. Como os gastos atuais são desconhecidos, são atribuídas as seguintes probabilidades, respectivamente: 0,3; 0,2; 0,25; 0,05; e 0,2.
    a. Mostre a distribuição de probabilidade correspondente à previsão de gastos.
    b. Qual é o valor esperado da previsão de gastos para o ano vindouro?
    c. Qual é a variância da previsão de gastos para o ano vindouro?
    d. Se as projeções de renda para o ano são estimadas em US$ 12 milhões, comente a posição financeira da universidade.
56. Uma pesquisa realizada pelo Bureau of Transportation Statistics (BTS) mostrou que o número médio de pessoas que usam meios de transporte gastam cerca de 26 minutos em uma viagem de um itinerário, de suas residências ao local de trabalho. Além disso, 5% das pessoas que usam meios de transporte relataram que fazem uma única viagem de um itinerário de mais de uma hora (http://www.bts.gov, 12 de janeiro de 2004).
    a. Se 20 pessoas que usam meios de transporte são entrevistadas em um dia em particular, qual é a probabilidade de três relatarem uma viagem de um itinerário de mais de uma hora?
    b. Se 20 pessoas que usam meios de transporte são entrevistadas em um dia em particular, qual é a probabilidade de nenhuma delas relatar uma viagem de um itinerário de mais de uma hora?
    c. Se uma empresa tem 2 mil funcionários, qual é o número esperado de funcionários que farão uma viagem com um itinerário de mais de uma hora?
    d. Se uma empresa tem 2 mil funcionários, qual é a variância e o desvio padrão do número de funcionários que fazem uma viagem de um itinerário de mais de uma hora?

57. Uma empresa planeja entrevistar usuários da internet para verificar como o seu site proposto será recebido por diferentes grupos etários. De acordo com o Census Bureau (Departamento do Censo), 40% das pessoas da faixa etária entre 18 e 54 anos e 12% das pessoas com 55 anos ou mais usam a internet (*Statistical Abstract of the United States*, 2000).
   a. Quantas pessoas da faixa etária entre 18 e 54 anos devem ser contatadas a fim de se descobrir o número esperado de pelo menos dez usuários da internet?
   b. Quantas pessoas da faixa etária a partir dos 55 anos devem ser contatadas para se ter o número esperado de pelo menos dez usuários da internet?
   c. Se você contatar o número de pessoas da faixa etária entre 18 e 54 anos sugerida no item (a), qual será o desvio padrão do número de pessoas que serão usuárias da internet?
   d. Se você contatar o número de pessoas da faixa etária a partir dos 55 anos sugerida no item (b), qual será o desvio padrão do número de pessoas que serão usuárias da internet?
58. Muitas empresas usam uma técnica de controle da qualidade denominada *amostragem de aceitação* para monitorar o carregamento de chegada de peças, matérias-primas e assim por diante. Na indústria eletrônica, os componentes comumente são despachados pelos fornecedores em grandes lotes. A inspeção de uma amostra de n componentes pode ser vista como os n ensaios de um experimento binomial. O resultado de cada componente testado (ensaio) indicará que ele é classificado como um componente bom ou defeituoso. A Reynolds Electronics aceita lotes de determinado fornecedor se os componentes defeituosos de um lote não ultrapassarem 1%. Suponha que uma amostra aleatória de cinco itens de uma remessa recente seja testada.
   a. Suponha que 1% da remessa apresente defeitos. Calcule a probabilidade de nenhum item da amostra estar defeituoso.
   b. Suponha que 1% da remessa apresente defeitos. Calcule a probabilidade de exatamente um item da amostra estar com defeito.
   c. Qual é a probabilidade de se observar um ou mais itens com defeito na amostra, se 1% da remessa tiver defeitos.
   d. Você se sentiria à vontade em aceitar a remessa se um item fosse considerado defeituoso? Por quê?
59. A taxa de desemprego é de 4,1% (*Barron's*, 4 de setembro de 2004). Suponha que 100 pessoas aptas a entrar no mercado de trabalho sejam selecionadas aleatoriamente.
   a. Qual é o número esperado de pessoas que estão desempregadas?
   b. Qual é a variância e o desvio padrão do número de pessoas que estão desempregadas?
60. Uma pesquisa de opinião levada a efeito pela Zogby International mostrou que, dos norte-americanos que disseram que a música desempenha papel "muito importante" em suas vidas, 30% disseram que as estações de rádio locais "sempre" executam o tipo de música de que eles gostam (http://www.zogby.com, 12 de janeiro de 2004). Suponha que seja tomada uma amostra de 800 pessoas que disseram que a música desempenha papel importante em suas vidas.
   a. Quantas pessoas você espera que digam que suas estações de rádio locais executam sempre o tipo de música de que elas gostam?
   b. Qual é o desvio padrão do número de entrevistados que acham que suas estações de rádio locais sempre executam o tipo de música de que eles gostam?
   c. Qual é o desvio padrão do número de entrevistados que não acha que suas estações de rádio locais sempre executam o tipo de música de que eles gostam?
61. Os carros chegam a um lava-rápido aleatória e independentemente; a probabilidade de um carro chegar é a mesma para dois intervalos de tempo de igual duração. A taxa média de chegada são 15 carros por hora. Qual é a probabilidade de 20 ou mais carros chegarem durante determinado horário de operação?
62. Um novo processo automatizado de produção tem uma média de 1,5 pane por dia. Em virtude do custo associado a cada pane, a administração está preocupada com a possibilidade de haver três ou mais panes durante um dia. Suponha que as panes ocorram aleatoriamente, que a probabilidade de uma pane seja a mesma para dois intervalos de tempo qualquer de igual duração e que as panes ocorridas em um período sejam independentes das panes ocorridas em outros períodos. Qual é a probabilidade de haver duas ou três panes durante um dia?
63. Um diretor regional responsável pelo desenvolvimento dos negócios no estado da Pensilvânia está preocupado com o número de fracassos de pequenos negócios. Se o número médio de fracassos de

pequenos negócios por mês for igual a 10, qual será a probabilidade de exatamente quatro pequenos negócios fracassarem durante determinado mês? Suponha que a probabilidade de fracassos seja a mesma para dois meses quaisquer e que a ocorrência ou não-ocorrência de um fracasso em determinado mês seja independente dos fracassos em outro mês qualquer.

64. Clientes chegam a um banco de forma aleatória e independente; a probabilidade de um cliente chegar no período de um minuto qualquer é similar à probabilidade de outro cliente chegar em outro período de um minuto qualquer. Responda às seguintes questões, supondo uma taxa de chegada média igual a três clientes por minuto.
    a. Qual é a probabilidade de exatamente três chegadas no período de um minuto?
    b. Qual é a probabilidade de haver pelo menos três chegadas no período de um minuto?
65. Um baralho contém 52 cartas, das quais quatro são ases. Qual é a probabilidade de uma mão de cinco cartas oferecer:
    a. Um par de ases?
    b. Exatamente um ás?
    c. Nenhum ás?
    d. Pelo menos um ás?
66. Durante a semana que se encerrou em 16 de setembro de 2001, Tiger Woods foi o vencedor que mais ganhou dinheiro no PGA Tour, com ganhos totais de US$ 5.517.777. Entre os dez principais vencedores, sete jogadores usaram uma bola de golfe marca Totleist (http://www.pgatour.com). Suponha que selecionemos aleatoriamente dois dos vencedores que mais ganharam dinheiro.
    a. Qual é a probabilidade de exatamente um usar a bola de golfe Titleist?
    b. Qual é a probabilidade de ambos usarem bolas de golfe Titleist?
    c. Qual é a probabilidade de nenhum deles usar uma bola de golfe Titleist?

Apêndice 5.1 – Distribuições Discretas de Probabilidade com o Minitab

Pacotes estatísticos como, por exemplo, o Minitab oferecem um procedimento relativamente eficiente e fácil para calcular probabilidades binomiais. Neste apêndice, ilustramos o procedimento etapa por etapa para se determinar as probabilidades binomiais relativas ao problema da loja de roupas do Martin apresentado na Seção 5.4. Lembre-se de que as probabilidades binomiais desejadas se baseiam em $n = 10$, e $p = 0{,}30$. Antes de iniciar a rotina do Minitab, o usuário deve inserir os valores desejados da variável aleatória x em uma coluna da planilha. Colocamos os valores 0, 1, 2, ..., 10 na coluna 1 (veja a Figura 5.5) para gerar toda a distribuição de probabilidade binomial. As etapas do Minitab para se obter as probabilidades binomiais desejadas são apresentadas a seguir.

Etapa 1. Selecione o menu **Calc**
Etapa 2. Escolha a opção **Probability Distributions**
Etapa 3. Escolha a opção **Binomial**
Etapa 4. Quando surgir a caixa de diálogo Distribuição Binomial:
Selecione **Probability**
Digite 10 na caixa **Number of trials**
Digite 0,3 na caixa **Probability of success**
Digite C1 na caixa **Input column**.
Dê um clique em **OK**

Os resultados do Minitab com as probabilidades binomiais terão a aparência mostrada na Figura 5.5.

O Minitab fornece as probabilidades de Poisson e hipergeométricas de maneira similar. Por exemplo, para calcular probabilidades de Poisson, as únicas diferenças estão na etapa 3, em que a opção **Poisson** deve ser selecionada, e na etapa 4, em que se deve digitar **Mean** em vez do número de ensaios e a probabilidade de sucesso.

Apêndice 5.2 – Distribuições Discretas de Probabilidade com o Excel

O Excel fornece funções para calcular probabilidades para as distribuições binomial, de Poisson e hipergeométrica, apresentadas neste capítulo. A função do Excel para calcular probabilidades binomiais é a BINOMDIST. Ela tem quatro argumentos: x (o número de sucessos), n (o número de ensaios), p (a proba-

bilidade de sucesso) e cumulativo. FALSE é usado para o quarto argumento (cumulativo), se quisermos a probabilidade de x sucessos, e TRUE é usada para o quarto argumento se quisermos a probabilidade cumulativa de x sucessos ou menos. Mostramos aqui como calcular as probabilidades de zero a dez sucessos para o problema da loja de roupa do Martin, mostrado na seção 5.4 (veja a Figura 5.5).

À medida que descrevermos o desenvolvimento da planilha, consulte a Figura 5.6; a planilha com a fórmula é definida em segundo plano e a planilha com o valor aparece em primeiro plano.

Figura 5.6 Planilha do Excel para calcular probabilidades binomiais

	A	B	C	D
1	**Número de Ensaios (n)**	10		
2	**Probabilidade de Sucesso (p)**	0.3		
3				
4		x	$f(x)$	
5		0	=BINOMDIST(B5,$B$1,$B$2,FALSE)	
6		1	=BINOMDIST(B6,$B$1,$B$2,FALSE)	
7		2	=BINOMDIST(B7,$B$1,$B$2,FALSE)	
8		3	=BINOMDIST(B8,$B$1,$B$2,FALSE)	
9		4	=BINOMDIST(B9,$B$1,$B$2,FALSE)	
10		5	=BINOMDIST(B10,$B$1,$B$2,FALSE)	
11		6	=BINOMDIST(B11,$B$1,$B$2,FALSE)	
12		7	=BINOMDIST(B12,$B$1,$B$2,FALSE)	
13		8	=BINOMDIST(B13,$B$1,$B$2,FALSE)	
14		9	=BINOMDIST(B14,$B$1,$B$2,FALSE)	
15		10	=BINOMDIST(B15,$B$1,$B$2,FALSE)	
16				

	A	B	C	D
1	**Número de Ensaios (n)**	10		
2	**Probabilidade de Sucesso (p)**	0.3		
3				
4		x	$f(x)$	
5		0	0,0282	
6		1	0,1211	
7		2	0,2335	
8		3	0,2668	
9		4	0,2001	
10		5	0,1029	
11		6	0,0368	
12		7	0,0090	
13		8	0,0014	
14		9	0,0001	
15		10	0,0000	
16				

Digitamos o número de ensaios (10) na célula B1, a probabilidade de sucesso na célula B2 e os valores da variável aleatória nas células B5:B15. As etapas seguintes gerarão as probabilidades desejadas:

Etapa 1. Use a função BINOMDIST para calcular a probabilidade de $x = 0$ ao digitar a seguinte fórmula na célula C5:

=BINOMDIST(B5,$B$1,$B$2,FALSE)

Etapa 2. Copie a fórmula da célula C5 para as células C6:C15.

A planilha de valor da Figura 5.6 mostra que as probabilidades obtidas são similares às apresentadas na Figura 5.5. As probabilidades de Poisson e hipergeométricas podem ser calculadas de maneira similar. São usadas as funções POISSON e HYPERGEOMETRIC. A ferramenta Insert Function do Excel pode ajudar o usuário a introduzir os argumentos necessários para estas funções (veja o Apêndice 2.2).

Distribuições Contínuas de Probabilidade

ESTATÍSTICA NA PRÁTICA

PROCTER & GAMBLE*
Cincinnati, Ohio

A Procter & Gamble (P&G) produz e comercializa produtos como detergentes, fraldas descartáveis, produtos farmacêuticos ao consumidor, cremes dentais, sabonetes, anti-sépticos bucais e toalhas de papel. Em nível mundial, sua marca ocupa a posição de liderança em mais categorias do que qualquer outra empresa de produtos de consumo.

Como líder na aplicação de métodos estatísticos para a tomada de decisões, a P&G emprega pessoas com os mais diversos tipos de formação acadêmica: engenharia, estatística, pesquisa operacional e administração. As principais tecnologias quantitativas para as quais esses profissionais dão suporte são: decisão probabilística e análise de riscos, simulação avançada, melhoria da qualidade e métodos quantitativos (por exemplo, programação linear, análise de regressão, análise de probabilidade).

A Industrial Chemicals Division da P&G é a principal fornecedora de alcoóis graxos derivados de substâncias naturais como o óleo de coco e derivados de petróleo. Essa divisão queria avaliar os riscos econômicos e as oportunidades de expandir suas instalações de produção de alcoóis graxos; portanto, foram convocados especialistas em decisão probabilística e análise de riscos da P&G para auxiliar. Depois de estruturar e esquematizar o problema, determinaram que a chave da lucratividade seria a diferença de custo entre as matérias-

* Os autores agradecem a Joel Kahn, da Procter & Gamble, por fornecer esta "Estatística na Prática".

primas à base de petróleo e de óleo de coco. Os custos futuros eram desconhecidos, mas os analistas puderam representá-los com as seguintes variáveis aleatórias contínuas.

x = o preço do óleo de coco por litro de álcool graxo.

e

y = o preço da matéria-prima à base de petróleo por quilo de álcool graxo.

Uma vez que a lucratividade era a diferença entre essas duas variáveis aleatórias, uma terceira variável aleatória, $d = x - y$, foi utilizada na análise. Especialistas foram entrevistados para determinar a distribuição de probabilidades de x e y.

Por sua vez, essa informação foi utilizada para desenvolverem uma distribuição contínua de probabilidade da diferença de preços d. Essa distribuição contínua de probabilidade forneceu a probabilidade de 0,90 de a diferença de preço ser de US$ 0,0655 ou menos, e a probabilidade de 0,50 de a diferença de preço ser de US$ 0,035 ou menos. Além disso, havia somente 0,10 de probabilidade de a diferença de preço ser de US$ 0,0045 ou menos.*

A Industrial Chemicals Division acreditava que o fato de serem capazes de quantificar o impacto das diferenças de preço das matérias-primas seria fundamental para chegar a um consenso. As probabilidades obtidas foram utilizadas em uma análise de sensibilidade da diferença de preços das matérias-primas. A análise produziu o *insight* suficiente para fundamentar uma recomendação à administração.

O uso de variáveis aleatórias contínuas e suas distribuições probabilísticas foi útil à P&G ao analisar os riscos econômicos associados à produção de alcoóis graxos. Neste capítulo, você compreenderá o que são as variáveis aleatórias contínuas e suas distribuições de probabilidade, incluindo uma das distribuições de probabilidade mais importantes da estatística: a distribuição normal.

No capítulo anterior, discutimos as variáveis aleatórias discretas e suas distribuições de probabilidade. Neste capítulo, voltamo-nos ao estudo das variáveis aleatórias contínuas. Especificamente, discutiremos três distribuições contínuas de probabilidade: a uniforme, a normal e a exponencial.

Uma diferença fundamental separa as variáveis aleatórias discretas e as contínuas em termos de como as probabilidades são calculadas. Quanto a uma variável aleatória discreta, a função de probabilidade $f(x)$ produz a probabilidade de a variável aleatória assumir um valor em particular. No que diz respeito às variáveis aleatórias contínuas, a contraparte da função de probabilidade é a **função densidade de probabilidade**, também expressa por $f(x)$. A diferença é que a função densidade de probabilidade não produz probabilidades diretamente. Entretanto, a área sob o gráfico de $f(x)$ correspondente a determinado intervalo produz a probabilidade de a variável aleatória contínua x assumir um valor nesse intervalo. Então, quando calculamos probabilidades de variáveis aleatórias contínuas, calculamos a probabilidade de a variável aleatória assumir qualquer valor nesse intervalo.

Uma das implicações da definição de probabilidade com respeito às variáveis aleatórias contínuas é o fato de a probabilidade de qualquer valor em particular da variável aleatória ser zero, porque a área sob o gráfico de $f(x)$ em qualquer ponto em particular é zero. Na Seção 6.1, demonstramos esses conceitos em relação a uma variável aleatória contínua que tem uma distribuição uniforme de probabilidade.

Grande parte deste capítulo dedica-se a descrever e ilustrar aplicações da distribuição normal de probabilidade. A distribuição normal de probabilidade tem importância fundamental em razão de sua ampla aplicabilidade e extenso uso na inferência estatística. Este capítulo encerra-se com uma discussão da distribuição exponencial de probabilidade.

---

* As diferenças de preço aqui apresentadas foram modificadas para guardar dados protegidos por direitos de propriedade.

6.1 DISTRIBUIÇÃO UNIFORME DE PROBABILIDADE

Suponha que a variável aleatória x represente o tempo de vôo de um avião que vai de Chicago a Nova York. Suponha que o tempo de vôo possa ter qualquer valor no intervalo de 120 a 140 minutos. Uma vez que a variável aleatória x pode assumir qualquer valor desse intervalo, x é uma variável aleatória contínua, não uma variável aleatória discreta. Suponha que suficientes dados de vôo reais estejam disponíveis para podermos concluir que a probabilidade de tempo de vôo no intervalo de 1 minuto qualquer tenha a mesma probabilidade de tempo de vôo em outro intervalo de 1 minuto contido no espaço mais amplo de 120 a 140 minutos. Considerando que cada um dos intervalos de 1 minuto é igualmente provável, dizemos que a variável aleatória tem uma **distribuição uniforme de probabilidade**. A função densidade de probabilidade, a qual define a distribuição uniforme de probabilidade correspondente à variável aleatória "tempo de vôo", é:

Sempre que a probabilidade é proporcional ao comprimento do intervalo, a variável aleatória se encontra uniformemente distribuída.

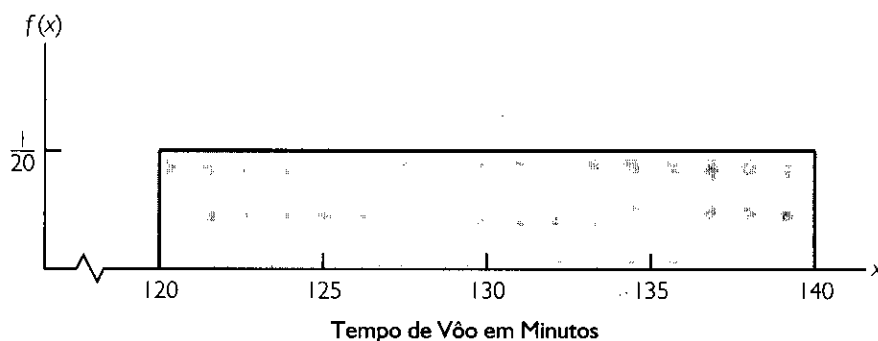
$$f(x) = \begin{cases} 1/20 & \text{para } 120 \leq x \leq 140 \\ 0 & \text{outro ponto qualquer} \end{cases}$$

A Figura 6.1 é um gráfico dessa função densidade de probabilidade. Geralmente, a função densidade uniforme de probabilidade de uma variável aleatória x é encontrada por meio da seguinte fórmula:

FUNÇÃO DENSIDADE UNIFORME DE PROBABILIDADE

$$f(x) = \begin{cases} \dfrac{1}{b-a} & \text{para } a \leq x \leq b \\ 0 & \text{outro ponto qualquer} \end{cases} \tag{6.1}$$

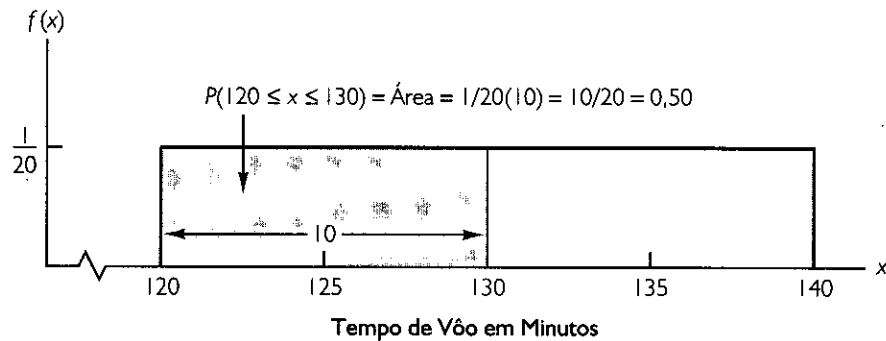
Em relação à variável aleatória "tempo de vôo", $a = 120$ e $b = 140$.

Figura 6.1 Função densidade uniforme da probabilidade de tempos de vôo

Conforme observamos na introdução com relação a uma variável aleatória contínua, consideramos a probabilidade somente em termos da possibilidade de uma variável aleatória assumir um valor dentro de um intervalo específico. No exemplo do tempo de vôo, uma questão de probabilidade aceitável é: qual é a probabilidade de o tempo de vôo situar-se entre 120 e 130 minutos? Ou seja, qual é $P(120 \leq x \leq 130)$? Visto que o tempo de vôo precisa estar entre 120 e 140 minutos, e porque a probabilidade é descrita como uniforme nesse intervalo, sentimo-nos à vontade para dizer que $P(120 \leq x \leq 130) = 0{,}50$. Na subseção seguinte, mostramos que essa probabilidade pode ser calculada como a área sob o gráfico de $f(x)$, de 120 a 130 (veja a Figura 6.2).

A Área como uma Medida de Probabilidade

Permita-nos fazer uma observação a respeito do gráfico da Figura 6.2. Considere a área sob o gráfico de $f(x)$ no intervalo entre 120 e 130. A área é retangular e sabemos que a área de um retângulo é simplesmen-

te a largura multiplicada pela altura. Sendo a largura do intervalo igual a 130 – 120 = 10, e a altura igual ao valor da função densidade de probabilidade $f(x) = 1/20$, temos a área, que é a largura multiplicada pela altura: 10(1/20) = 10/20 = 0,50.

Figura 6.2 A área fornece a probabilidade do tempo de vôo entre 120 e 130 minutos

Qual observação você poderia fazer a respeito da área sob o gráfico de $f(x)$ e a probabilidade? Elas são idênticas! De fato, essa observação é verdadeira para todas as variáveis aleatórias contínuas. Tão logo a função densidade de probabilidade $f(x)$ seja identificada, a probabilidade de x assumir um valor entre algum valor x_1 mais baixo e algum valor x_2 mais alto pode ser encontrada calculando-se a área sob o gráfico de $f(x)$ no intervalo entre x_1 e x_2.

Dada a distribuição uniforme do tempo de vôo, e usando a área como uma probabilidade, podemos responder a quaisquer questões probabilísticas sobre os tempos de vôo. Por exemplo, qual é a probabilidade de ocorrência de um tempo de vôo entre 128 e 136 minutos? A largura do intervalo é 136 – 128 = 8. Sendo a altura de $f(x) = 1/20$ uniforme, observamos que $P(128 \leq x \leq 136) = 8(1/20) = 0{,}40$.

Observe que $P(120 \leq x \leq 140) = 20(1/20) = 1$, ou seja, a área total sob o gráfico de $f(x)$ é igual a 1. Essa propriedade é válida para todas as distribuições contínuas de probabilidade e é análoga à condição de que a soma das probabilidades deve ser igual a 1 em uma função de probabilidade discreta. No que se refere a uma função densidade contínua de probabilidade, também devemos impor que $f(x) \geq 0$ para todos os valores de x. Esse requisito é análogo à necessidade de se ter $f(x) \geq 0$ para funções de probabilidade discretas.

Para constatar a veracidade de que a probabilidade de um ponto simples qualquer é 0, consulte a Figura 6.2 e calcule a probabilidade de um ponto simples, digamos, $x = 125$. $P(x = 125) = P(125 \leq x \leq 125) = 0(1/20) = 0$.

Duas importantes diferenças se colocam no tratamento das variáveis aleatórias contínuas e no tratamento de suas contrapartes discretas.

1. Não falamos mais da probabilidade de a variável aleatória assumir um valor em particular. Ao contrário, falamos da probabilidade de a variável aleatória assumir um valor dentro de um intervalo determinado.
2. A probabilidade de uma variável aleatória contínua assumir um valor dentro de determinado intervalo entre x_1 e x_2 é definida como a área sob o gráfico da função densidade de probabilidade que se encontra entre x_1 e x_2. Uma vez que um ponto simples é um intervalo que tem largura zero, isso implica que a probabilidade de uma variável aleatória contínua assumir de maneira exata qualquer valor em particular é zero. Significa também que a probabilidade de uma variável aleatória contínua assumir um valor em qualquer intervalo é a mesma, quer os pontos extremos sejam incluídos quer não.

O cálculo do valor esperado e da variância de uma variável aleatória contínua é análogo ao cálculo que efetuamos para uma variável aleatória discreta. Entretanto, desde que o procedimento de cálculo envolva cálculo integral, deixamos a derivação das fórmulas apropriadas para os livros mais avançados.

Quanto à distribuição contínua uniforme de probabilidade introduzida nesta seção, as fórmulas do valor esperado e da variância são as seguintes:

$$E(x) = \frac{a + b}{2}$$

$$\text{Var}(x) = \frac{(b - a)^2}{12}$$

Nessas fórmulas, a é o menor valor, e b, o maior valor que a variável aleatória pode assumir.

Aplicando essas fórmulas à distribuição uniforme de probabilidade para os tempos de vôo Chicago a Nova York, obtemos

$$E(x) = \frac{(120 + 140)}{2} = 130$$

$$\text{Var}(x) = \frac{(140 - 120)^2}{12} = 33{,}33$$

O desvio padrão dos tempos de vôo pode ser encontrado extraindo-se a raiz quadrada da variância. Desse modo, $\sigma = 5{,}77$ minutos.

NOTAS E COMENTÁRIOS

Para entender com mais clareza por que a altura de uma função densidade de probabilidade não é uma probabilidade, imagine uma variável aleatória com a seguinte distribuição uniforme de probabilidade:

$$f(x) = \begin{cases} 2 & \text{para } 0 \leq x \leq 0{,}5 \\ 0 & \text{outro ponto qualquer} \end{cases}$$

A altura da função densidade de probabilidade, $f(x)$, é 2 para os valores de x situados entre 0 e 0,5. Porém, sabemos que as probabilidades nunca podem ser maiores que 1. Desse modo, notamos que $f(x)$ não pode ser interpretada como a probabilidade de x.

Exercícios

Métodos

AUTOTESTE

1. Sabe-se que a variável aleatória x está distribuída uniformemente entre 1,0 e 1,5.
   a. Apresente o gráfico da função densidade de probabilidade.
   b. Calcule $P(x = 1{,}25)$.
   c. Calcule $P(1{,}0 \leq x \leq 1{,}25)$.
   d. Calcule $P(1{,}20 < x < 1{,}5)$.
2. Sabe-se que a variável aleatória x está distribuída uniformemente entre 10 e 20.
   a. Apresente o gráfico da função densidade de probabilidade.
   b. Calcule $P(x < 15)$.
   c. Calcule $P(12 \leq x \leq 18)$.
   d. Calcule $E(x)$.
   e. Calcule $\text{Var}(x)$.

Aplicações

3. A Delta Airlines declara que seus tempos de vôo de Cincinnati a Tampa são de duas horas e cinco minutos. Suponha que acreditemos que os tempos de vôo reais estejam uniformemente distribuídos no intervalo de duas horas e duas horas e 20 minutos.
   a. Apresente o gráfico da função densidade de probabilidade correspondente aos tempos de vôo.
   b. Qual é a probabilidade de o vôo ter não mais que cinco minutos de atraso?
   c. Qual é a probabilidade de o vôo ter mais que dez minutos de atraso?
   d. Qual é a expectativa do tempo de vôo?

AUTOTESTE

4. A maioria das linguagens de computador contém uma função que pode ser usada para gerar números aleatórios. No Excel, a função ALEATÓRIO pode ser usada para gerar números aleatórios entre 0 e 1. Se admitirmos que x denota um número aleatório gerado pela função ALEATÓRIO, então x é uma variável aleatória contínua com a seguinte função densidade de probabilidade:

$$f(x) = \begin{cases} 1 & \text{para } 0 \leq x \leq 1 \\ 0 & \text{outro ponto qualquer} \end{cases}$$

a. Trace o gráfico da função densidade de probabilidade.
b. Qual é a probabilidade de se gerar um número aleatório entre 0,25 e 0,75?
c. Qual é a probabilidade de se gerar um número aleatório com valor menor ou igual a 0,30?
d. Qual é a probabilidade de se gerar um número aleatório com valor maior que 0,60?

5. A maior distância de arremesso obtida pelos cem melhores golfistas da PGA (*Professional Golfers Association*) situa-se entre 284,7 e 310,6 jardas (*Golfweek*, 29 de março de 2003). Suponha que a maior distância de arremesso obtida por esses golfistas se distribua uniformemente ao longo desse intervalo.
   a. Apresente uma expressão matemática da função densidade de probabilidade da maior distância de arremesso.
   b. Qual é a probabilidade de a maior distância de arremesso obtida por um desses golfistas ser menor que 290 jardas?
   c. Qual é a probabilidade de a maior distância de arremesso obtida por um desses golfistas ser de, no mínimo, 300 jardas?
   d. Qual é a probabilidade de a maior distância de arremesso de um desses golfistas se situar entre 290 e 305 jardas?
   e. Quantos desses golfistas arremessam a bola, no mínimo, 290 jardas?
6. O rótulo de uma garrafa de detergente líquido indica que o conteúdo é de 12 onças por garrafa. A operação de produção preenche a garrafa uniformemente, de acordo com a seguinte função densidade de probabilidade:

$$f(x) = \begin{cases} 8 & \text{para } 11{,}975 \leq x \leq 12{,}100 \\ 0 & \text{outro ponto qualquer} \end{cases}$$

   a. Qual é a probabilidade de uma garrafa ser preenchida com um volume entre 12 e 12,05 onças?
   b. Qual é a probabilidade de uma garrafa ser preenchida com 12,02 onças ou mais?
   c. O controle da qualidade aceita uma margem de erro de 0,02 onças no preenchimento de uma garrafa em relação ao volume indicado em seu rótulo. Qual é a probabilidade de a garrafa desse detergente líquido deixar de cumprir o padrão estabelecido pelo controle da qualidade?
7. Suponha que estejamos interessados em apresentar uma oferta de compra de um lote de terra e sabemos que há outro concorrente interessado.[1] O vendedor anunciou que a oferta mais alta, acima de US$ 10 mil, seria aceita. Suponha que a oferta x apresentada pelo concorrente seja uma variável aleatória que se distribui uniformemente entre US$ 10 mil e US$ 15 mil.
   a. Suponha que você faça uma oferta de US$ 12 mil. Qual é a probabilidade de o seu lance ser aceito?
   b. Suponha que você faça uma oferta de US$ 14 mil. Qual é a probabilidade de o seu lance ser aceito?
   c. Qual valor você deve oferecer para maximizar a probabilidade de obter a propriedade?
   d. Suponha que você conheça alguém que esteja disposto a pagar US$ 16 mil pela propriedade. Você consideraria fazer uma oferta menor que o valor envolvido no item (c)? Por quê?

6.2 DISTRIBUIÇÃO NORMAL DE PROBABILIDADE

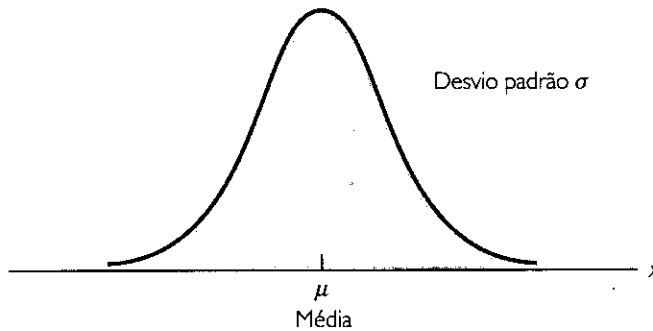
Abraham de Moivre, matemático francês, publicou *The Doctrine of Chances* em 1733. Foi ele quem deduziu a distribuição normal de probabilidade.

A mais importante distribuição de probabilidade para descrever uma variável aleatória contínua é a **distribuição normal de probabilidade**. A distribuição normal de probabilidade é usada em ampla variedade de aplicações práticas em que as variáveis aleatórias são a altura e peso das pessoas, notas de exames, medições científicas, índices pluviométricos e outros valores similares. Ela também é amplamente usada na inferência estatística, a qual corresponde o tópico principal do restante deste livro. Nessas aplicações, a distribuição normal fornece uma descrição dos resultados prováveis obtidos por meio de amostragem.

[1] Esse exercício baseia-se em um problema sugerido pelo professor Roger Myerson, da Northwestern University.

Curva Normal

O formato, ou forma, da distribuição normal de probabilidade é ilustrado pela curva em forma de sino apresentada na Figura 6.3. A função densidade de probabilidade que define a curva em forma de sino da distribuição normal de probabilidade é a seguinte:

Figura 6.3 Curva em forma de sino correspondente à distribuição normal de probabilidade

FUNÇÃO DENSIDADE NORMAL DE PROBABILIDADE

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2} \tag{6.2}$$

em que

m = média
s = desvio padrão
p = 3,14159
e = 2,7182

Vamos fazer diversas observações sobre as carcterísticas da distribuição normal.

A curva normal tem dois parâmetros, μ e σ. Eles determinam a posição e a forma da distribuição normal de probabilidade.

1. A família inteira das distribuições normais de probabilidade é diferenciada por dois parâmetros: sua média m e seu desvio padrão s.
2. O ponto máximo da curva normal encontra-se na média, que é também a mediana e a moda da distribuição.
3. A média da distribuição pode ser qualquer valor numérico: negativo, zero ou positivo. Três distribuições normais com o mesmo desvio padrão, mas três diferentes médias, (–10, 0 e 20), são mostradas a seguir:

4. A distribuição normal é simétrica, sendo a forma da curva à esquerda da média uma imagem espelhada da forma da curva à direita da média. Os extremos (caudas) da curva tendem ao infinito em ambas as direções e, teoricamente, jamais tocam o eixo horizontal. Uma vez que é simétrica, a distribuição normal de probabilidade não é inclinada; a medida de sua assimetria é zero.

5. O desvio padrão determina quanto uma curva é achatada ou larga. Valores maiores do desvio padrão resultam em curvas mais largas e mais achatadas, exibindo maior variabilidade dos dados. Duas distribuições normais com a mesma média, mas com desvios-padrão diferentes são apresentadas a seguir:

6. As probabilidades da variável aleatória normal são dadas por áreas sob a curva. A área total sob a curva correspondente à distribuição normal é 1. Já que a distribuição é simétrica, a área sob a curva, à esquerda da média, é 0,50, e a área sob a curva, à direita da média, é 0,50.
7. As porcentagens dos valores de alguns intervalos comumente usados são:
   a. 68,3% dos valores de uma variável aleatória normal estão dentro de mais ou menos um desvio padrão de sua média.
   b. 95,4% dos valores de uma variável aleatória normal estão dentro de mais ou menos dois desvios padrão de sua média.
   c. 99,7% dos valores de uma variável aleatória normal estão dentro de mais ou menos três desvios padrão de sua média.

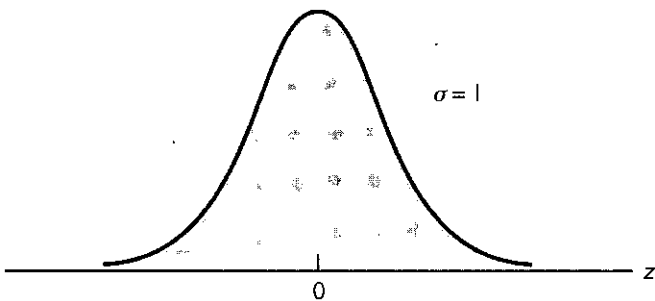
Essas porcentagens são a base para a regra empírica introduzida na Seção 3.3.

A Figura 6.4 apresenta graficamente as propriedades (a), (b) e (c).

Distribuição Normal Padrão de Probabilidade

Dizemos que a variável aleatória que tem uma distribuição normal cuja média é zero e o desvio padrão 1 tem uma **distribuição normal padrão de probabilidade**. Comumente, usamos a letra z para designar essa variável aleatória normal em particular. A Figura 6.5 representa o gráfico da distribuição normal padrão. Ela tem a mesma aparência geral das outras distribuições normais, porém com as propriedades especiais de $\mu = 0$ e $\sigma = 1$.

Figura 6.4 As áreas sob a curva de uma distribuição normal qualquer

Figura 6.5 A distribuição normal padrão

Uma vez que $\mu = 0$ e $\sigma = 1$, a fórmula da função densidade normal padrão de probabilidade é uma versão mais simples da Equação 6.2.

FUNÇÃO DENSIDADE NORMAL PADRÃO DE PROBABILIDADE

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2}$$

À semelhança de outras variáveis aleatórias contínuas, os cálculos de probabilidade com quaisquer distribuições normais são feitos calculando-se as áreas sob o gráfico da função densidade de probabilidade. Desse modo, para encontrar a probabilidade de uma variável aleatória normal estar dentro de um intervalo específico, devemos calcular a área sob a curva normal ao longo desse intervalo. Quanto à distribuição normal padrão, as áreas sob a curva normal foram calculadas e estão disponíveis em tabelas que podem ser usadas no cálculo das probabilidades. A Tabela 6.1 é uma delas, a qual também está disponível com o título de Tabela 1 no Apêndice B e na parte interna da primeira capa deste livro.

Quanto à função densidade normal de probabilidade, a altura da curva varia e são necessários cálculos matemáticos mais avançados para calcular as áreas que representam a probabilidade.

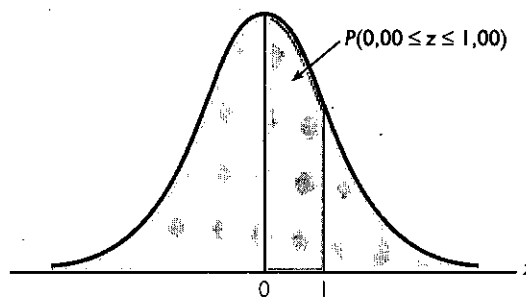
Para ver como se pode usar a tabela de áreas sob a curva da distribuição normal padrão (Tabela 6.1) para encontrar probabilidades, vamos considerar alguns exemplos. Posteriormente, veremos como essa mesma tabela pode ser usada para calcular as probabilidades de qualquer distribuição normal.

Tabela 6.1 As áreas, ou probabilidades, da distribuição normal padrão

z	0,00	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
0,0	0,0000	0,0040	0,0080	0,0120	0,0160	0,0199	0,0239	0,0279	0,0319	0,0359
0,1	0,0398	0,0438	0,0478	0,0517	0,0557	0,0596	0,0636	0,0675	0,0714	0,0753
0,2	0,0793	0,0832	0,0871	0,0910	0,0948	0,0987	0,1026	0,1064	0,1103	0,1141
0,3	0,1179	0,1217	0,1255	0,1293	0,1331	0,1368	0,1406	0,1443	0,1480	0,1517
0,4	0,1554	0,1591	0,1628	0,1664	0,1700	0,1736	0,1772	0,1808	0,1844	0,1879
0,5	0,1915	0,1950	0,1985	0,2019	0,2054	0,2088	0,2123	0,2157	0,2190	0,2224
0,6	0,2257	0,2291	0,2324	0,2357	0,2389	0,2422	0,2454	0,2486	0,2517	0,2549
0,7	0,2580	0,2611	0,2642	0,2673	0,2704	0,2734	0,2764	0,2794	0,2823	0,2852
0,8	0,2881	0,2910	0,2939	0,2967	0,2995	0,3023	0,3051	0,3078	0,3106	0,3133
0,9	0,3159	0,3186	0,3212	0,3238	0,3264	0,3289	0,3315	0,3340	0,3365	0,3389

Tabela 6.1 As áreas, ou probabilidades, da distribuição normal padrão (continuação)

1,0	0,3413	0,3438	0,3461	0,3485	0,3508	0,3531	0,3554	0,3577	0,3599	0,3621
1,1	0,3643	0,3665	0,3686	0,3708	0,3729	0,3749	0,3770	0,3790	0,3810	0,3830
1,2	0,3849	0,3869	0,3888	0,3907	0,3925	0,3944	0,3962	0,3980	0,3997	0,4015
1,3	0,4032	0,4049	0,4066	0,4082	0,4099	0,4115	0,4131	0,4147	0,4162	0,4177
1,4	0,4192	0,4207	0,4222	0,4236	0,4251	0,4265	0,4279	0,4292	0,4306	0,4319
1,5	0,4332	0,4345	0,4357	0,4370	0,4382	0,4394	0,4406	0,4418	0,4429	0,4441
1,6	0,4452	0,4463	0,4474	0,4484	0,4495	0,4505	0,4515	0,4525	0,4535	0,4545
1,7	0,4554	0,4564	0,4573	0,4582	0,4591	0,4599	0,4608	0,4616	0,4625	0,4633
1,8	0,4641	0,4649	0,4656	0,4664	0,4671	0,4678	0,4686	0,4693	0,4699	0,4706
1,9	0,4713	0,4719	0,4726	0,4732	0,4738	0,4744	0,4750	0,4756	0,4761	0,4767
2,0	0,4772	0,4778	0,4783	0,4788	0,4793	0,4798	0,4803	0,4808	0,4812	0,4817
2,1	0,4821	0,4826	0,4830	0,4834	0,4838	0,4842	0,4846	0,4850	0,4854	0,4857
2,2	0,4861	0,4864	0,4868	0,4871	0,4875	0,4878	0,4881	0,4884	0,4887	0,4890
2,3	0,4893	0,4896	0,4898	0,4901	0,4904	0,4906	0,4909	0,4911	0,4913	0,4916
2,4	0,4918	0,4920	0,4922	0,4925	0,4927	0,4929	0,4931	0,4932	0,4934	0,4936
2,5	0,4938	0,4940	0,4941	0,4943	0,4945	0,4946	0,4948	0,4949	0,4951	0,4952
2,6	0,4953	0,4955	0,4956	0,4957	0,4959	0,4960	0,4961	0,4962	0,4963	0,4964
2,7	0,4965	0,4966	0,4967	0,4968	0,4969	0,4970	0,4971	0,4972	0,4973	0,4974
2,8	0,4974	0,4975	0,4976	0,4977	0,4977	0,4978	0,4979	0,4979	0,4980	0,4981
2,9	0,4981	0,4982	0,4982	0,4983	0,4984	0,4984	0,4985	0,4985	0,4986	0,4986
3,0	0,4987	0,4987	0,4987	0,4988	0,4988	0,4989	0,4989	0,4989	0,4990	0,4990

Para começar, vejamos como podemos calcular a probabilidade de o valor z correspondente à variável aleatória normal padrão estar entre 0,00 e 1,00; ou seja, $P(0,00 \leq z \leq 1,00)$. A região sombreada com uma cor mais escura no gráfico a seguir exibe essa probabilidade.

Os lançamentos feitos na Tabela 6.1 fornecem a área sob a curva normal padrão entre a média $z = 0$ e um valor específico de z (veja o gráfico na parte superior da tabela). Nesse caso, estamos interessados na área entre $z = 0$ e $z = 1,00$. Então, precisamos encontrar na tabela o lançamento que corresponde a $z = 1,00$. Primeiramente, localizamos 1,0 na coluna à esquerda da tabela e depois encontramos 0,00 em sua linha superior. Examinando o corpo da tabela, descobrimos que a linha 1,0 e a coluna 0,00 se interceptam no valor 0,3413, o qual nos dá a probabilidade desejada: $P(0,00 \leq z \leq 1,00) = 0,3413$. Apresentamos, a seguir, uma parte da Tabela 6.1, a qual nos mostra estas etapas:

z	0,00	0,01	0,02
.			
.			
.			
0,9	0,3159	0,3186	0,3212
1,0	0,3413	0,3438	0,3461
1,1	0,3643	0,3665	0,3686
1,2	0,3849	0,3869	0,3888
.			
.			
.			

$P(0,00 \leq z \leq 1,00)$

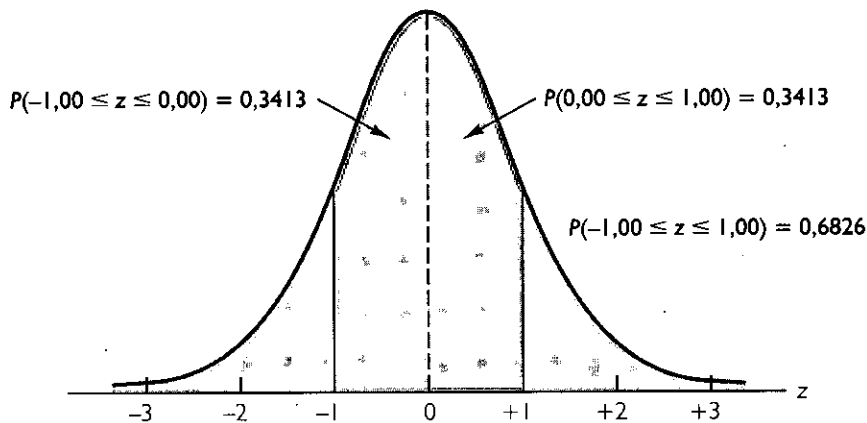
Usando a mesma abordagem, podemos encontrar $P(0,00 \leq z \leq 1,25)$. Localizando primeiramente a linha 1,2 e deslocando-nos lateralmente na tabela até a coluna 0,05, encontramos $P(0,00 \leq z \leq 1,25) =$ 0,3944.

Como outro exemplo do uso da tabela de áreas da distribuição normal padrão, calculamos a probabilidade de obtermos um valor $z = -1,00$ e $z = 1,00$; ou seja, $P(-1,00 \leq z \leq 1,00)$.

Note que já usamos a Tabela 6.1 para mostrar que a probabilidade de haver um valor z entre $z = 0,00$ e $z = 1,00$ é 0,3413, e lembre-se de que a distribuição normal é *simétrica*. Desse modo, a probabilidade de haver um valor z entre $z = 0,00$ e $z = -1,00$ é idêntica à probabilidade de haver um valor z entre $z =$ 0,00 e $z = +1,00$. Portanto, a probabilidade de haver um valor z entre $z = -00$ e $z = +1,00$ é:

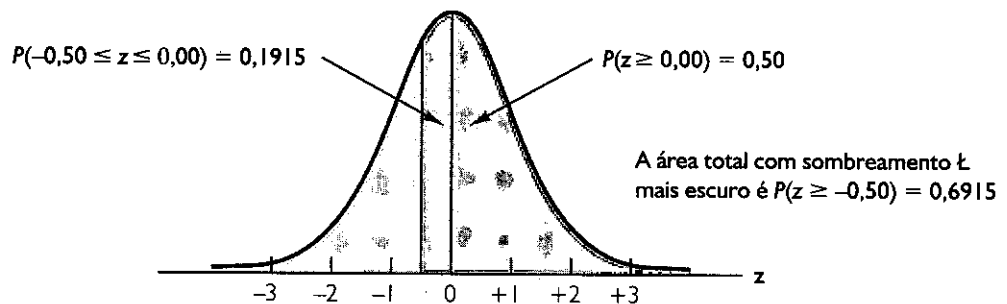
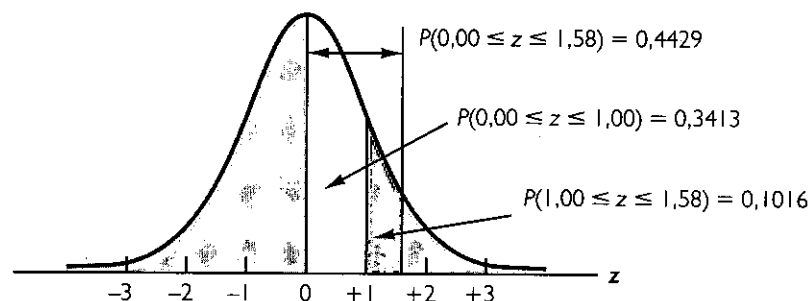
$$P(-1,00 \leq z \leq 0,00) + P(0,00 \leq z \leq 1,00) = 0,3413 + 0,3413 = 0,6826$$

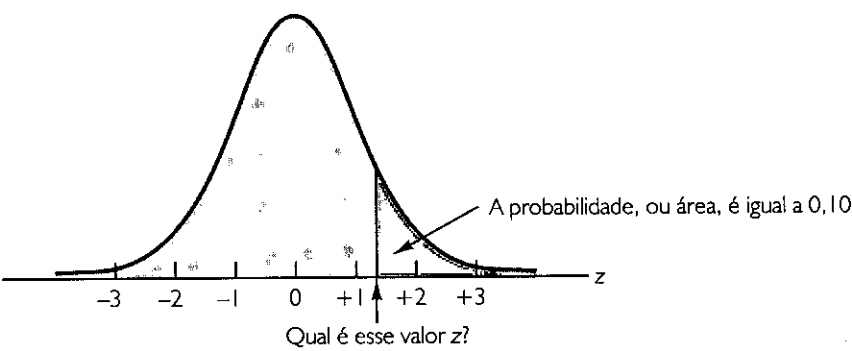
Essa probabilidade é apresentada graficamente na figura a seguir:

Esses cálculos de probabilidade são a base para a observação 7 apresentada na página 212.

De maneira similar, podemos usar os valores da Tabela 6.1 para demonstrar que a probabilidade de haver um valor z entre $-2,00$ e $+2,00$ é $0,4772 + 0,4772 = 0,9544$, e que a probabilidade de haver um valor z entre $-3,00$ e $+3,00$ é $0,4987 + 0,4987 = 0,9974$. Já que sabemos que a probabilidade total – ou a área total sob a curva de qualquer variável aleatória contínua – deve ser 1,0000, a probabilidade 0,9974 nos diz que o valor de z quase sempre estará entre $-3,00$ e $+3,00$.

Calculamos a seguir a probabilidade de obtermos um valor z de, no mínimo, 1,58; ou seja, $P(z \geq 1,58)$. Primeiramente, usamos a linha $z = 1,5$ e a coluna 0,08 da Tabela 6.1, e descobrimos que $P(0,00 \leq z \leq 1,58)$ $= 0,4429$. Ora, como a distribuição normal de probabilidade é simétrica, sabemos que 50% da área sob a curva devem estar à direita da média (isto é, $z = 0$) e 50% da área sob a curva devem estar à esquerda da média. Se 0,4429 é a área entre a média e $z = 1,58$, então a área, ou probabilidade, correspondente a $z \geq$ 1,58 deve ser $0,5000 - 0,4429 = 0,0571$. Essa probabilidade é apresentada na figura a seguir:

Como outro exemplo, considere a probabilidade de a variável aleatória z assumir o valor –0,50 ou maior; ou seja, $P(z \geq -0{,}50)$. Para fazermos esse cálculo, observamos que a probabilidade que procuramos pode ser escrita como a soma de duas probabilidades: $P(z \geq 0{,}50) = P(-0{,}50 \leq z \leq 0{,}00) + P(z \geq 0{,}00)$. Vimos anteriormente que $P(z \geq 0{,}00) = 0{,}50$. Além disso, sabemos também que, desde que a distribuição normal seja simétrica, $P(-0{,}50 \leq z \leq 0{,}00) = P(0{,}00 \leq z \leq 0{,}50)$.

Consultando a Tabela 6.1, descobrimos que $P(0{,}00 \leq z \leq 0{,}50) = 0{,}1915$. Portanto, $P(z \geq -0{,}50) = 0{,}1915 + 0{,}5000 = 0{,}6915$. O gráfico a seguir mostra essa probabilidade:

Calculamos a seguir a probabilidade de obtermos um valor z entre 1,00 e 1,58; ou seja, $P(1{,}00 \leq z \leq 1{,}58)$. De nossos exemplos anteriores, sabemos que há 0,3413 de probabilidade de um valor z estar entre $z = 0{,}00$ e $z = 1{,}00$, e que há 0,4429 de probabilidade de um valor z estar entre $z = 0{,}00$ e $z = 1{,}58$. Portanto, deve haver uma probabilidade $0{,}4429 - 0{,}3413 = 0{,}1016$ de um valor z estar entre $z = 1{,}00$ e $z = 1{,}58$. Desse modo, $P(1{,}00 \leq z \leq 1{,}58) = 0{,}1016$. Essa situação é mostrada graficamente na figura a seguir:

Como ilustração final, encontremos um valor z tal que a probabilidade de obtermos um valor z mais elevado seja 0,10. A figura seguinte apresenta essa situação graficamente:

Dada uma probabilidade, podemos usar a tabela normal padrão de modo inverso para encontrar o valor z correspondente.

Esse cálculo é o inverso daquele que usamos nos exemplos anteriores. Anteriormente, especificamos o valor z de interesse e depois encontramos a probabilidade, ou área, correspondente. Nesse exemplo, fornecemos a probabilidade, ou área, e pedimos que se encontre o valor z correspondente. Para fazê-lo, usamos a tabela de probabilidades da distribuição normal padrão (Tabela 6.1) de uma maneira bem diferente.

Lembre-se de que o corpo da Tabela 6.1 fornece a área sob a curva existente entre a média e um valor de z em particular. Possuímos a informação de que a área na extremidade (cauda) superior da curva é 0,10. Portanto, precisamos determinar quanto da área está entre a média e o valor z de interesse. Como sabemos que 0,5000 da área está à direita da média, 0,5000 – 0,1000 = 0,4000 deve ser a área sob a curva existente entre a média e o valor z desejado. Fazendo uma varredura no corpo da tabela, encontramos 0,3997 como o valor probabilístico mais próximo de 0,4000. Apresentamos a seguir a parte da tabela que fornece esse resultado.

z	0,06	0,07	0,08	0,09
.				
.				
.				
1,0	0,3554	0,3577	0,3599	0,3621
1,1	0,3770	0,3790	0,3810	0,3830
1,2	0,3962	0,3980	0,3997	0,4015
1,3	0,4131	0,4147	0,4162	0,4177
1,4	0,4279	0,4292	0,4306	0,4319
.				
.				

Valor da área mais próximo de 0,4000, no corpo da tabela

Verificando o valor z na coluna da extrema esquerda e na linha do topo da tabela, descobrimos que o valor z correspondente é 1,28. Desse modo, uma área de aproximadamente 0,4000 (0,3997, de fato) estará entre a média e z = 1,28.[2] Em termos da pergunta formulada originalmente, a probabilidade é de aproximadamente 0,10 de que o valor z seja maior que 1,28.

[2] Poderíamos usar interpolação no corpo da tabela para obtermos uma aproximação melhor do valor de z correspondente à área de 0,4000. Isso nos garantiria a precisão de mais uma casa decimal e produziria um valor z igual a 1,282. Entretanto, na maioria das situações práticas, a precisão suficiente é obtida simplesmente usando-se os valores da tabela mais próximos da probabilidade desejada.

Os exemplos ilustram que a tabela de áreas da distribuição normal padrão pode ser usada para se encontrar probabilidades associadas a valores da variável aleatória normal padrão z.

Dois tipos de questão podem ser apresentados. O primeiro tipo especifica um valor, ou valores, de z e nos pede para usarmos a tabela para determinar as áreas, ou probabilidades, correspondentes. O segundo fornece uma área, ou probabilidade, e nos pede para usarmos a tabela para determinar os valores z correspondentes. Assim, precisamos ser flexíveis ao usar a tabela normal padrão para responder à questão de probabilidade desejada. Na maioria dos casos, esboçar um gráfico da distribuição normal padrão e sombrear a área, ou probabilidade, apropriada, ajuda a visualizar a situação e auxilia na determinação da resposta correta.

Como Calcular Probabilidades de Qualquer Distribuição Normal

A fórmula da variável aleatória normal padrão é similar à que introduzimos no Capítulo 3 para calcular contagens-z de um conjunto de dados.

A razão para discutirmos tão extensamente a distribuição normal padrão é que as probabilidades de todas as distribuições normais são calculadas usando-se a distribuição normal padrão. Ou seja, quando temos uma distribuição normal com uma média μ qualquer e um desvio padrão σ qualquer, respondemos às questões de probabilidade referentes à distribuição efetuando primeiramente a conversão para distribuição normal padrão. Então, podemos usar a Tabela 6.1 e os valores apropriados z para encontrar as probabilidades desejadas. A fórmula usada para converter qualquer variável aleatória normal x com média μ e desvio padrão σ em distribuição normal padrão é apresentada a seguir:

COMO CONVERTER EM DISTRIBUIÇÃO NORMAL PADRÃO

$$z = \frac{x - \mu}{\sigma} \tag{6.3}$$

Um valor de x igual à sua média μ resulta em $z = (\mu - \mu)/\sigma = 0$. Desse modo, vemos que um valor de x igual à sua média μ corresponde a um valor de z em sua média 0. Suponha agora que x seja um desvio padrão maior que sua média; ou seja, $x = \mu + \sigma$. Aplicando a Equação 6.3, notamos que o valor z correspondente é $z = [(\mu + \sigma) - \mu]/\sigma = \sigma/\sigma = 1$. Assim, um valor de x que está um desvio padrão acima de sua média corresponde a $z = 1$. Em outras palavras, podemos interpretar z como o número de desvios padrão que a variável aleatória normal x está afastada de sua média μ.

Para ver como essa conversão nos possibilita calcular as probabilidades de qualquer distribuição normal, suponha que tenhamos uma distribuição normal com $\mu = 10$ e $\sigma = 2$. Qual é a probabilidade de a variável aleatória x estar entre 10 e 14? Usando a Equação 6.3, notamos que para $x = 10$, $z = (x - \mu)/\sigma = (10 - 10)/2 = 0$ e que para $x = 14$, $z = (14 - 10)/2 = 4/2 = 2$. Então, a resposta para a nossa questão sobre a probabilidade de x estar entre 10 e 14 é dada pela probabilidade equivalente de z estar entre 0 e 2 em relação à distribuição normal padrão. Em outras palavras, a probabilidade que procuramos é a probabilidade de a variável aleatória x estar entre sua média e dois desvios padrão acima da média. Usando $z = 2{,}00$ e a Tabela 6.1, observamos que a probabilidade é 0,4772. Por isso, a probabilidade de x estar entre 10 e 14 é 0,4772.

O Problema da Grear Tire Company

Voltamo-nos agora a uma aplicação da distribuição normal. Suponha que a Grear Tire Company tenha desenvolvido um novo pneu radial com cinturão de aço que será vendido por meio de uma cadeia nacional de *discount stores*.[3] Uma vez que esse tipo de pneu é um novo produto, os gerentes da Grear acreditam que a durabilidade (em termos de milhas rodadas) oferecida com o pneu será um fator importante na aceitação do produto. Antes de fechar os termos do contrato de garantia de durabilidade do pneu, os gerentes da Grear desejam obter informações de probabilidade a respeito do número de milhas que os pneus durarão.

Dos testes reais de estrada com os pneus, a equipe de engenharia da Grear estima que a durabilidade média dos pneus é $\mu = 36.500$ milhas (58.741 quilômetros) e que o desvio padrão é $\sigma = 5.000$. Além disso, os dados coletados indicam que a distribuição normal é uma hipótese razoável.

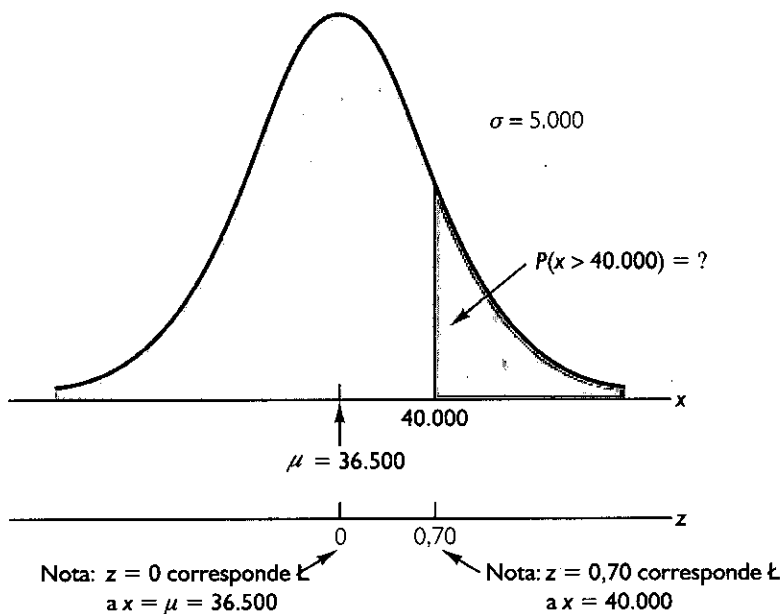
Qual porcentagem dos pneus possivelmente duraria mais de 40 mil milhas (64.373 quilômetros)? Em outras palavras, qual é a probabilidade de a durabilidade do pneu ultrapassar 40 mil milhas? Essa questão pode ser respondida encontrando-se a área da região com sombreamento mais forte na Figura 6.6.

[3] NT: *Discount store* – Literalmente, "loja de descontos". Estabelecimento comercial (geralmente, de cadeias de lojas) que vende seus produtos por preços mais baixos.

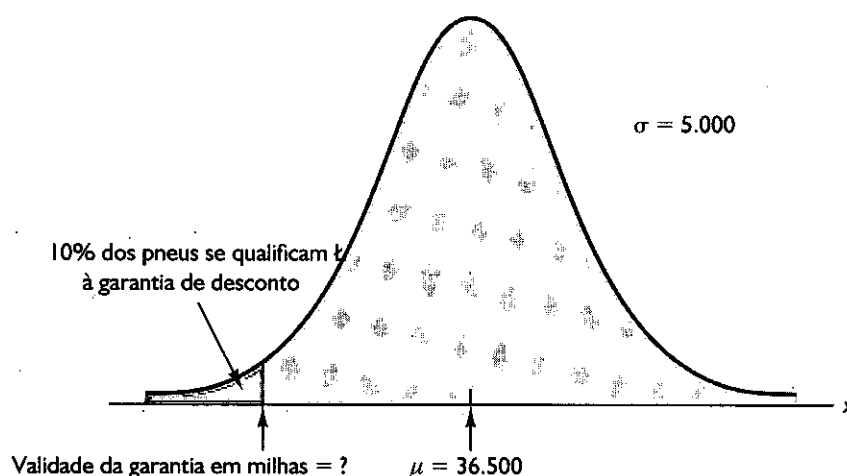
Para $x = 40.000$, temos

$$z = \frac{x-\mu}{\sigma} = \frac{40.000 - 36.500}{5.000} = \frac{3.500}{5.000} = 0,70$$

Consultemos agora a parte inferior da Figura 6.6. Notamos que um valor $x = 40.000$ na distribuição normal da Grear Tire corresponde a um valor $z = 0,70$ na distribuição normal padrão. Usando a Tabela 6.1, observamos que a área entre a média e $z = 0,70$ é 0,2580. Consultando novamente a Figura 6.6, observamos que a área entre $x = 36.500$ e $x = 40.000$ na distribuição normal da Grear Tire também é a mesma (0,2580). Desse modo, $0,5000 - 0,2580 = 0,2420$ é a probabilidade de x ultrapassar 40.000. Podemos concluir que aproximadamente 24,2% dos pneus terão uma durabilidade maior que 40 mil milhas.

Figura 6.6 Distribuição da durabilidade dos pneus da Grear Tire Company em termos de milhas

Suponhamos agora que a Grear esteja considerando a possibilidade de dar uma garantia que concede um desconto na troca de pneus se os originais não resistirem ao número de milhas estipulado na garantia. Qual deve ser o número de milhas coberto pela garantia levando-se em conta que a Grear quer que não mais de 10% dos pneus se habilitem à garantia do desconto? Essa questão é interpretada graficamente na Figura 6.7.

De acordo com a Figura 6.7, 40% da área deve estar entre a média e o número de milhas desconhecido a ser coberto pela garantia.

Figura 6.7 Garantia de desconto da Grear Tire Company

Procuramos 0,4000 no corpo da Tabela 6.1. Por simetria, a área procurada está aproximadamente 1,28 desvios padrão à esquerda da média. Ou seja, $z = -1{,}28$ é o valor da variável aleatória normal padrão correspondente à validade da garantia desejada em termos de milhas na distribuição normal da Grear Tire Company. Para encontrar o valor de x correspondente a $z = -1{,}28$, calculamos

$$z = \frac{x - \mu}{\sigma} = -1{,}28$$

$$x - \mu = -1{,}28\sigma$$

$$x = \mu - 1{,}28\sigma$$

Sendo $\mu = 36.500$ e $\sigma = 5.000$,

$$x = 36.500 - 1{,}28(5.000) = 30.100$$

Com a garantia fixada em 30 mil milhas (48.280 km), a porcentagem real apta à garantia será de 9,68%.

Assim, uma garantia de 30.100 milhas (48.280 km) cumprirá o requisito de que aproximadamente 10% dos pneus se habilitem à garantia. Talvez, com essa informação, a empresa possa fixar a garantia de durabilidade de seus pneus em 30 mil milhas.

Novamente, constatamos o importante papel que as distribuições de probabilidade desempenham em termos de produzir informações para a tomada de decisões. Ou seja, assim que uma distribuição de probabilidade é estabelecida para uma aplicação em particular, ela pode ser usada rápida e facilmente para se obter informações a respeito do problema. A probabilidade não determina a recomendação de uma decisão diretamente, mas fornece informações que ajudam o tomador de decisão a entender melhor os riscos e as incertezas associados ao problema. Por fim, essas informações podem auxiliá-lo a tomar uma boa decisão.

Exercícios

Métodos

8. Usando a Figura 6.4 como guia, esboce a curva normal de uma variável aleatória x que tem a média $\mu = 100$ e desvio padrão $\sigma = 10$. Rotule o eixo horizontal com valores 70, 80, 90, 100, 110, 120 e 130.
9. Uma variável aleatória normalmente se distribui com uma média de $\mu = 50$ e um desvio padrão de $\sigma = 5$.
   a. Esboce uma curva normal da função densidade de probabilidade. Rotule o eixo horizontal com os valores 35, 40, 45, 50, 55, 60 e 65. A Figura 6.4 mostra que a curva normal quase toca o eixo horizontal em três desvios padrão abaixo e em três desvios padrão acima da média (nesse caso, em 35 e 65).
   b. Qual é a probabilidade de a variável aleatória assumir um valor entre 45 e 55?
   c. Qual é a probabilidade de a variável aleatória assumir um valor entre 40 e 60?

10. Trace um gráfico da distribuição normal padrão. Rotule o eixo horizontal nos valores –3, –2, –1, 0, 1, 2 e 3. Depois use a tabela de probabilidades da distribuição normal padrão para calcular as seguintes probabilidades:
    a. $P(0 \leq z \leq 1)$.
    b. $P(0 \leq z \leq 1{,}5)$.
    c. $P(0 < z < 2)$.
    d. $P(0 < z < 2{,}5)$.
11. Dado que z é uma variável aleatória normal padrão, calcule as seguintes probabilidades:
    a. $P(-1 \leq z \leq 0)$.
    b. $P(-1{,}5 \leq z \leq 0)$.
    c. $P(-2 < z < 0)$.
    d. $P(-2{,}5 \leq z \leq 0)$.
    e. $P(-3 < z \leq 0)$.
12. Dado que z é uma variável aleatória normal padrão, calcule as seguintes probabilidades:
    a. $P(0 \leq z \leq 0{,}83)$.
    b. $P(-1{,}57 \leq z \leq 0)$.
    c. $P(z > 0{,}44)$.
    d. $P(z \geq -0{,}23)$.
    e. $P(z < 1{,}20)$.
    f. $P(z \leq -0{,}71)$.

AUTOTESTE

13. Dado que z é uma variável aleatória normal padrão, calcule as seguintes probabilidades:
    a. $P(-1{,}98 \leq z \leq 0{,}49)$.
    b. $P(0{,}52 \leq z \leq 1{,}22)$.
    c. $P(-1{,}75 \leq z \leq -1{,}04)$.
14. Dado que z é uma variável aleatória normal padrão, encontre z para cada uma das situações:
    a. A área entre 0 e z é 0,4750.
    b. A área entre 0 e z é 0,2291.
    c. A área à direita de z é 0,1314.
    d. A área à esquerda de z é 0,6700.
15. Dado que z é uma variável aleatória normal padrão, encontre z para cada uma das situações:
    a. A área à esquerda de z é 0,2119.
    b. A área entre $-z$ e z é 0,9030.
    c. A área entre $-z$ e z é 0,2052.
    d. A área à esquerda de z é 0,9948.
    e. A área à direita de z é 0,6915.

AUTOTESTE

16. Dado que z é uma variável aleatória normal, encontre z para cada uma das situações:
    a. A área à direita de z é 0,01.
    b. A área à direita de z é 0,025.
    c. A área à direita de z é 0,05.
    d. A área à direita de z é 0,10.

Aplicações

17. A quantia média que pais e filhos gastaram por criança na compra de roupas para o retorno às aulas no outono de 2001 foi de US$ 527 (*CNBC*, 5 de setembro de 2001). Suponha que o desvio padrão seja US$ 160 e que a quantia gasta esteja distribuída normalmente.
    a. Qual é a probabilidade de a quantia gasta com uma criança escolhida aleatoriamente ser superior a US$ 700?
    b. Qual é a probabilidade de a quantia gasta com uma criança escolhida aleatoriamente ser inferior a US$ 100?
    c. Qual é a probabilidade de a quantia gasta com uma criança escolhida aleatoriamente estar entre US$ 450 e US$ 700?
    d. Qual é a probabilidade de a quantia gasta com uma criança escolhida aleatoriamente não ultrapassar US$ 300?

AUTOTESTE

18. A média de preço das ações das empresas que compõem a S&P 500 é US$ 30, e o desvio padrão é US$ 8,20 (*Business Week*, edição especial anual, primavera de 2003). Suponha que os preços das ações se distribuam normalmente.
    a. Qual é a probabilidade de uma empresa ter um preço de, no mínimo, US$ 40 para suas ações?
    b. Qual é a probabilidade de uma empresa ter um preço não superior a US$ 20 para suas ações?
    c. Qual deve ser o preço das ações para que a empresa seja incluída entre as 10% maiores?
19. A média pluviométrica durante o mês de abril em Dallas, Texas, é de 88,9 milímetros (*The World Almanac*, 2000). Suponha que uma distribuição normal seja aplicável e que o desvio padrão seja de 20,32 mm.
    a. Em qual porcentagem do tempo a quantidade de chuva ultrapassou 127 mm em abril?
    b. Em qual porcentagem do tempo a quantidade de chuva foi inferior a 76,2 mm em abril?
    c. Um mês é classificado como extremamente úmido se a quantidade de chuva se situar nos 10% superior em relação a esse mês. Quanta chuva deve cair para que um mês de abril seja classificado como extremamente úmido?
20. Em janeiro de 2003 o trabalhador norte-americano passou em média 77 horas conectado à internet enquanto se encontrava no trabalho (*CNBC*, 15 de março de 2003). Suponha que os tempos estejam normalmente distribuídos e que o desvio padrão seja de 20 horas.
    a. Qual é a probabilidade de um trabalhador escolhido aleatoriamente passar menos de 50 horas conectado à internet?
    b. Qual porcentagem de trabalhadores passaram mais de 100 horas conectados à internet?
    c. Uma pessoa é classificada como forte usuário se estiver entre os 20% que fazem mais uso. Quantas horas um trabalhador deve manter-se conectado à internet para ser classificado como forte usuário?
21. Uma pessoa deve obter uma pontuação entre os 2% mais bem classificados da população em um teste de QI para afiliar-se à Mensa, uma sociedade internacional de pessoas com QI elevado (*US Airways Attache*, setembro de 2000). Se as pontuações de QI forem normalmente distribuídas com uma média 100 e desvio padrão igual a 15, qual pontuação uma pessoa deve obter para poder afiliar-se à Mensa?
22. De acordo com o Bureau of Labor Statistics, a remuneração média por semana dos trabalhadores norte-americanos do setor de produção foi de US$ 441,84 (*The World Almanac*, 2000). Suponha que os dados disponíveis indiquem que os salários dos trabalhadores do setor de produção estejam normalmente distribuídos, com um desvio padrão de US$ 90.
    a. Qual é a probabilidade de um trabalhador ter ganho um salário entre US$ 400 e US$ 500?
    b. Quanto um trabalhador do setor de produção teve de ganhar para se colocar entre os 20% que receberam os maiores salários?
    c. Em relação a um trabalhador do setor de produção escolhido aleatoriamente, qual é a probabilidade de ele ter ganho menos de US$ 250 por semana?
23. O tempo necessário para concluir um exame final em determinado curso universitário está distribuído normalmente com uma média de 80 minutos e desvio padrão de dez minutos. Responda às seguintes questões:
    a. Qual é a probabilidade de alguém concluir o exame em uma hora ou menos?
    b. Qual é a probabilidade de um estudante concluir o exame em mais de 60 minutos, porém, menos de 75 minutos?
    c. Suponha que a classe tenha 60 alunos e que a duração do exame seja de 90 minutos. Quantos estudantes você acha que não conseguirão concluir o exame no tempo determinado?
24. O volume diário (milhões de ações) de títulos negociados na Bolsa de Valores de Nova York durante 12 dias de agosto e setembro é mostrado a seguir (*Barron's*, 7 de agosto de 2000, 4 de setembro de 2000 e 11 de setembro de 2000).

917	983	1.046
944	723	783
813	1.057	766
836	992	973

A distribuição de probabilidade do volume de negócios é aproximadamente normal.

a. Calcule a média e o desvio padrão do volume diário de negócios para usá-los como estimativas da média da população e do desvio padrão.
b. Qual é a probabilidade de, em determinado dia, o volume de negócios ser inferior a 800 milhões de ações?
c. Qual é a probabilidade de o volume de negócios ultrapassar um bilhão de ações?
d. Se a Bolsa de Valores quiser emitir um *release* sobre os 5% melhores dias de negócios, qual volume motivará um *release*?

25. O preço médio dos ingressos para um jogo de futebol do Washington Redskins na temporada de 2001 foi de US$ 81,89 (*USA Today*, 6 de setembro de 2001). Com os custos adicionais de estacionamento, alimentação, bebidas e *souvenirs*, o custo médio para uma família de quatro pessoas assistir ao jogo totalizava US$ 442,54. Suponha que se aplique a distribuição normal e que o desvio padrão seja US$ 65,00.
   a. Qual é a probabilidade de uma família de quatro pessoas gastar mais de US$ 400,00?
   b. Qual é a probabilidade de uma família de quatro pessoas gastar US$ 300,00 ou menos?
   c. Qual é a probabilidade de uma família de quatro pessoas gastar entre US$ 400,00 e US$ 500,00?

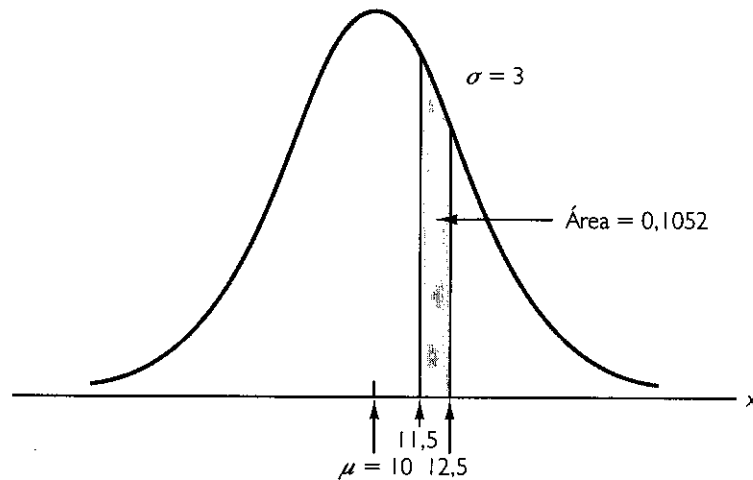
6.3 APROXIMAÇÃO NORMAL ÀS PROBABILIDADES BINOMIAIS

Na Seção 5.4, apresentamos a distribuição binomial de probabilidade discreta. Lembre-se de que um experimento binomial consiste em uma seqüência de n ensaios independentes e idênticos, tendo cada ensaio dois resultados possíveis: um sucesso ou um fracasso. A probabilidade de um sucesso em um ensaio é a mesma para todos os ensaios e é denotada por p. A variável aleatória binomial é o número de sucessos obtidos nos n ensaios e as questões probabilísticas dizem respeito à probabilidade de x sucessos nos n ensaios.

Quando o número de ensaios torna-se grande, é difícil calcular a função binomial de probabilidade manualmente ou com o auxílio de uma calculadora. Nos casos em que $np \geq 5$ e $n(1-p) \geq 5$, a distribuição normal fornece uma aproximação fácil de usar às probabilidades binomiais. Quando usamos a aproximação normal à probabilidade binomial, ajustamos $\mu = np$ e $\sqrt{np(1-p)}$ na definição da curva normal.

Vamos ilustrar a aproximação normal à probabilidade binomial supondo que uma empresa privada tem em seu histórico o fato de cometer erros em 10% de suas faturas. Foi tomada uma amostra de cem faturas, e queremos calcular a probabilidade de 12 faturas conterem erros. Ou seja, queremos encontrar a probabilidade binomial de 12 sucessos em cem ensaios. Ao aplicar a aproximação normal nesse caso, determinamos que $\mu = np = (100)(0{,}10) = 10$ e $\sqrt{np(1-p)} = \sqrt{(100)(0{,}1)(0{,}9)} = 3$. Uma distribuição normal com $\mu = 10$ e $\sigma = 3$ é mostrada na Figura 6.8.

Lembre-se de que, quando se trata de uma distribuição contínua de probabilidade, as probabilidades são calculadas como áreas sob a função densidade de probabilidade. Conseqüentemente, a probabilidade de um valor único qualquer para a variável aleatória é zero. Desse modo, para fazermos a aproximação à probabilidade binomial de 12 sucessos, calculamos a área sob a curva normal correspondente, entre 11,5 e 12,5. O 0,5 que adicionamos e subtraímos de 12 é chamado **fator de correção de continuidade**. Ele é introduzido porque utilizamos uma distribuição contínua para aproximar uma distribuição discreta. Então, o $P(x = 12)$ da distribuição binomial *discreta* é aproximado por $P(11{,}5 \leq x \leq 12{,}5)$, da distribuição normal *contínua*.

Figura 6.8 Aproximação normal a uma distribuição binomial de probabilidade, com $n = 100$ e $p = 0,10$ mostrando a probabilidade de 12 erros

Efetuando a conversão para a distribuição normal padrão para calcularmos $P(11,5 \leq x \leq 12,5)$, obtemos:

$$z = \frac{x - \mu}{\sigma} = \frac{12,5 - 10,0}{3} = 0,83 \text{ para } x = 12,5$$

e

$$z = \frac{x - \mu}{\sigma} = \frac{11,5 - 10,0}{3} = 0,50 \text{ para } x = 11,5$$

Na Tabela 6.1, descobrimos que a área sob a curva (na Figura 6.8) entre 10 e 12,5 é 0,2967. Analogamente, a área sob a curva entre 10 e 11,5 é 0,1915. Portanto, a área entre 11,5 e 12,5 é 0,2967 – 0,1915 = 0,1052. A aproximação normal à probabilidade de 12 sucessos em 100 ensaios é 0,1052.

Como outro exemplo, suponha que queiramos calcular a probabilidade de 13 erros ou menos em uma amostra de 100 faturas. A Figura 6.9 mostra a área sob a curva normal que faz a aproximação a essa probabilidade. Observe que o uso do fator de correção de continuidade tem como conseqüência o fato de o valor 13,5 ser usado para calcular a probabilidade desejada. O valor z correspondente a $x = 13,5$ é:

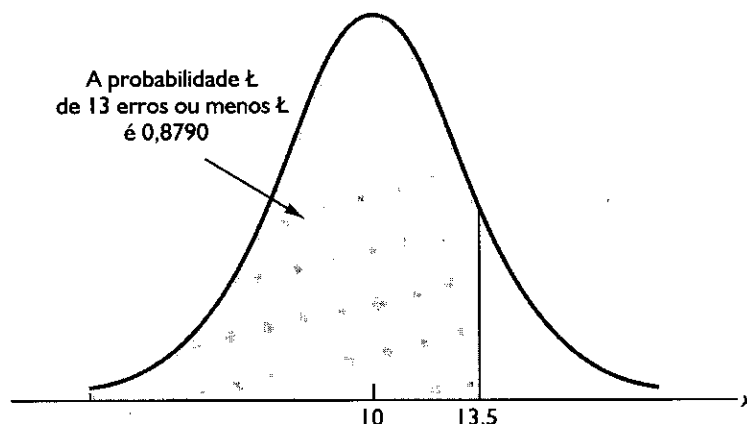
$$z = \frac{13,5 - 10,0}{3,0} = 1,17$$

A Tabela 6.1 mostra que a área sob a curva normal padrão entre 0 e 1,17 é 0,3790. A área sob a curva normal que faz a aproximação à probabilidade de 13 erros ou menos é dada pela parte sombreada do gráfico apresentado na Figura 6.9. A probabilidade é 0,3790 + 0,5000 = 0,8790.

Figura 6.9 Aproximação normal a uma distribuição binomial de probabilidade, com $n = 100$ e $p = 0{,}10$ mostrando a probabilidade de 13 erros ou menos

Exercícios

Métodos

AUTOTESTE

26. Uma distribuição binomial de probabilidade tem $p = 0{,}20$ e $n = 100$.
    a. Qual é a média e qual é o desvio padrão?
    b. Essa é uma daquelas situações em que as probabilidades binomiais podem ser aproximadas pela distribuição normal de probabilidade? Explique.
    c. Qual é a probabilidade de haver exatamente 24 sucessos?
    d. Qual é a probabilidade de 18 a 22 sucessos?
    e. Qual é a probabilidade de 15 sucessos ou menos?
27. Suponha que uma distribuição binomial de probabilidade tem $p = 0{,}60$ e $n = 200$.
    a. Qual é a média e qual é o desvio padrão?
    b. Essa é uma daquelas situações em que as probabilidades binomiais podem ser aproximadas pela distribuição normal de probabilidade? Explique.
    c. Qual é a probabilidade de 100 a 110 sucessos?
    d. Qual é a probabilidade de 130 sucessos ou mais?
    e. Qual é a vantagem de usarmos a distribuição normal de probabilidade para aproximar as probabilidades binomiais? Use o item (d) para explicar a vantagem.

Aplicações

AUTOTESTE

28. O presidente Bush propôs a eliminação dos impostos sobre os dividendos pagos aos acionistas sob a alegação de que eles resultam em dupla tributação. Os rendimentos usados para pagar os dividendos já são tributados às corporações. Uma pesquisa sobre essa questão revelou que 47% dos norte-americanos são favoráveis à proposta. Por partido político, 64% dos republicanos e 20% dos democratas são favoráveis à proposta (*Investor's Business Daily*, 13 de janeiro de 2003). Suponha que um grupo de 250 norte-americanos se reúna para ouvir uma palestra sobre a proposta.
    a. Qual é a probabilidade de pelo menos a metade do grupo ser favorável à proposta?
    b. Você descobre depois que 150 republicanos e 100 democratas estão presentes. Agora, qual é a sua estimativa do número esperado de pessoas que são favoráveis à proposta?
    c. Um orador favorável à proposta será mais bem recebido por esse grupo do que alguém contrário à proposta?
29. A taxa de desemprego é 5,8% (*Bureau of Labor Statistics*, www.bls.gov, 3 de abril de 2003). Suponha que cem pessoas aptas ao trabalho sejam selecionadas aleatoriamente.
    a. Qual é o número esperado de pessoas desempregadas?
    b. Qual é a variância e o desvio padrão do número de desempregados?
    c. Qual é a probabilidade de exatamente seis estarem desempregados?
    d. Qual é a probabilidade de pelo menos quatro estarem desempregados?

30. Ao assinar um contrato de cartão de crédito você o lê cuidadosamente? Em uma pesquisa realizada pela FindLaw.com, foi feita a seguinte pergunta às pessoas: "Quão minuciosamente você lê um contrato de cartão de crédito?" (*USA Today*, 6 de outubro de 2003). As descobertas revelaram que 44% lêem cada palavra, 33% lêem o suficiente para entender o contrato, 11% dão apenas uma olhada e 4% não o lêem absolutamente.
    a. Em relação a uma amostra de 500 pessoas, quantas você acha que diriam que lêem cada palavra de um contrato de cartão de crédito?
    b. Em relação a uma amostra de 500 pessoas, qual é a probabilidade de 200 ou menos dizerem que lêem cada palavra de um contrato de cartão de crédito?
    c. Em relação a uma amostra de 500 pessoas, qual é a probabilidade de pelo menos 15 dizerem que não lêem os contratos de cartão de crédito?
31. Um hotel da estância turística de Myrtle Beach tem 120 quartos. Nos meses de primavera, a ocupação dos quartos do hotel é de aproximadamente 75%.
    a. Qual é a probabilidade de pelo menos metade dos quartos estarem ocupados em determinado dia?
    b. Qual é a probabilidade de 100 ou mais quartos estarem ocupados em determinado dia?
    c. Qual é a probabilidade de 80 ou menos quartos estarem ocupados em determinado dia?

6.4 DISTRIBUIÇÃO EXPONENCIAL DE PROBABILIDADE

A **distribuição exponencial de probabilidade** pode ser usada para variáveis aleatória, como os intervalos de tempo de chegada dos carros a um lava-rápido, o tempo necessário para carregar um caminhão, a distância entre defeitos importantes em uma rodovia e assim por diante. A função densidade exponencial de probabilidade é apresentada a seguir:

FUNÇÃO DENSIDADE EXPONENCIAL DE PROBABILIDADE

$$f(x) = \frac{1}{\mu} e^{-x/\mu} \quad \text{para } x \geq 0, \mu > 0 \qquad \textbf{(6.4)}$$

Como um exemplo da distribuição exponencial, suponha que x represente o tempo de carga de um caminhão no terminal de carga da Schips e que ele siga esse tipo de distribuição. Se o valor médio, ou a média, do tempo de carga for 15 minutos ($\mu = 15$), a função densidade de probabilidade apropriada será:

$$f(x) = \frac{1}{15} e^{-x/15}$$

A Figura 6.10 é o gráfico dessa função densidade de probabilidade.

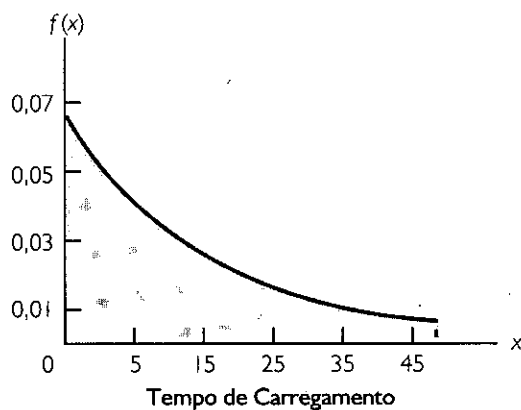
Como Calcular Probabilidades da Distribuição Exponencial

Em aplicações de fila de espera, a distribuição exponencial freqüentemente é usada para o tempo de atendimento.

À semelhança do que ocorre com qualquer distribuição contínua de probabilidade, a área sob a curva correspondente a um intervalo fornece a probabilidade de a variável aleatória assumir um valor nesse intervalo. No exemplo do terminal de carga da Schips, a probabilidade de o carregamento de um caminhão demandar seis minutos ou menos ($x \leq 6$) é definida como a área sob a curva representada na Figura 6.10, de $x = 0$ a $x = 6$. Similarmente, a probabilidade de o tempo de carregamento de um caminhão demandar 18 minutos ou menos ($x \leq 18$) é a área sob a curva, de $x = 0$ a $x = 18$.

Observe também que a probabilidade de o tempo de carregamento de um caminhão se situar entre seis e 18 minutos ($6 \leq x \leq 18$) é dada pela área sob a curva, de $x = 6$ a $x = 18$.

Para calcular probabilidades exponenciais como as que acabamos de descrever, usamos a fórmula apresentada a seguir. Ela fornece a probabilidade cumulativa de obtermos um valor menor ou igual a um valor específico de x, denotado por x_0, para a variável aleatória exponencial.

Figura 6.10 Distribuição exponencial de probabilidade referente ao exemplo do terminal de carga da Schips

DISTRIBUIÇÃO EXPONENCIAL: PROBABILIDADES CUMULATIVAS

$$P(x \leq x_0) = 1 - e^{-x_0/\mu} \qquad (6.5)$$

Em relação ao exemplo do terminal de carga da Schips, x = tempo de carregamento e $\mu = 15$, o que nos dá:

$$P(x \leq x_0) = 1 - e^{-x_0/15}$$

Portanto, a probabilidade de o carregamento de um caminhão demandar seis minutos ou menos é:

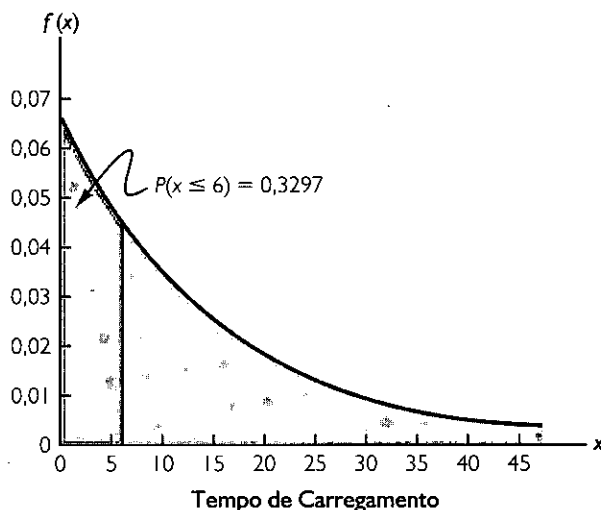
$$P(x \leq 6) = 1 - e^{-6/15} = 0{,}3297$$

A Figura 6.11 apresenta a área, ou a probabilidade, de um tempo de carregamento de seis minutos ou menos.

Usando a Equação 6.5, calculamos a probabilidade de se carregar um caminhão em 18 minutos ou menos.

$$P(x \leq 18) = 1 - e^{-18/15} = 0{,}6988$$

Desse modo, a probabilidade de o tempo de carregamento do um caminhão demandar entre seis e 18 minutos é igual a 0,6988 – 0,3297 = 0,3691. As probabilidades correspondentes a qualquer outro intervalo podem ser calculadas de maneira similar.

Figura 6.11 Probabilidade de ocorrer um tempo de carregamento igual a seis minutos ou menos

Uma propriedade da distribuição exponencial é que a média e o desvio padrão são iguais.

No exemplo anterior, o tempo médio necessário para carregar um caminhão é $\mu = 15$ minutos. Uma propriedade da distribuição exponencial é que tanto a média quanto o desvio padrão da distribuição são *iguais*. Assim, o desvio padrão do tempo necessário para carregar um caminhão é $s = 15$ minutos. A variância é $\sigma^2 = (15)^2 = 225$.

Relações entre a Distribuição de Poisson e a Distribuição Exponencial

Na Seção 5.5 introduzimos a distribuição de Poisson como uma distribuição de probabilidade discreta que muitas vezes é útil para examinarmos o número de ocorrências de um evento ao longo de um intervalo específico de tempo ou de espaço. Lembre-se de que a função de probabilidade de Poisson é:

$$f(x) = \frac{\mu^x e^{-\mu}}{x!}$$

em que

μ = o valor esperado, ou número médio, de ocorrências ao longo de um intervalo específico

Se as chegadas seguem uma distribuição de Poisson, o tempo entre as chegadas deve seguir uma distribuição exponencial.

A distribuição exponencial contínua de probabilidade está relacionada à distribuição discreta de Poisson. Se a distribuição de Poisson fornece uma descrição apropriada do número de ocorrências por intervalo, a distribuição exponencial fornece uma descrição da extensão do intervalo entre as ocorrências.

Para ilustrar essa relação, suponha que o número de carros que chegam a um lava-rápido durante uma hora seja descrito por uma distribuição de probabilidade de Poisson, com uma média de dez carros por hora. A função de probabilidade de Poisson que dá a probabilidade de x chegadas por hora é:

$$f(x) = \frac{10^x e^{-10}}{x!}$$

Uma vez que o número médio de chegadas é de dez carros por hora, o tempo médio entre os carros que chegam é:

$$1 \text{ hora} = \frac{0{,}1 \text{ hora/carro}}{10 \text{ carros}}$$

Desse modo, a distribuição exponencial correspondente que descreve o tempo entre as chegadas tem uma média de $\mu = 0{,}1$ hora por carro; em conseqüência, a função densidade exponencial de probabilidade apropriada é:

$$f(x) = \frac{1}{0{,}1} e^{-x/0{,}1} = 10e^{-10x}$$

NOTAS E COMENTÁRIOS

Como podemos observar na Figura 6.10, a distribuição exponencial tem uma inflexão à direita. De fato, a medida de assimetria das distribuições exponenciais é 2. A distribuição exponencial nos dá uma boa idéia de como se apresenta uma distribuição assimétrica.

Exercícios

Métodos

32. Considere a seguinte função densidade exponencial de probabilidade.

$$f(x) = \frac{1}{8} e^{-x/8} \quad \text{para } x \geq 0$$

a. Encontre $P(x \leq 6)$.
b. Encontre $P(x \leq 4)$.

c. Encontre $P(x \geq 6)$.
d. Encontre $P(4 \leq x \leq 6)$.

33. Considere a seguinte função densidade exponencial de probabilidade.

$$f(x) = \frac{1}{3} e^{-x/3} \quad \text{para } x \geq 0$$

AUTOTESTE

a. Escreva a fórmula para $P(x \leq x_0)$.
b. Encontre $P(x \leq 2)$.
c. Encontre $P(x \geq 3)$.
d. Encontre $P(x \leq 5)$.
e. Encontre $P(2 \leq x \leq 5)$.

Aplicações

34. A *Internet Magazine* monitora provedores de internet (ISPs) e divulga estatísticas sobre seu desempenho. O tempo médio para fazer *download* (baixar) de uma página da rede de provedores de acesso gratuito é aproximadamente 20 segundos quando se trata de páginas da Web européias (*Internet Magazine*, janeiro de 2000). Suponha que o tempo para baixar uma página da internet siga uma distribuição exponencial.

a. Qual é a probabilidade de ser necessário menos de 10 segundos para baixar uma página da Web?
b. Qual é a probabilidade de ser necessário mais de 20 segundos para baixar uma página da Web?
c. Qual é a probabilidade de ser necessário entre 10 e 30 segundos para baixar uma página da Web?

AUTOTESTE

35. O tempo entre a chegada dos veículos a determinado cruzamento segue uma distribuição exponencial de probabilidade, com uma média de 12 segundos.

a. Apresente um esboço dessa distribuição exponencial de probabilidade.
b. Qual é a probabilidade de o tempo de chegada entre os veículos ser de 12 segundos ou menos?
c. Qual é a probabilidade de o tempo de chegada entre os veículos ser de 6 segundos ou menos?
d. Qual é a probabilidade de transcorrer 30 segundos ou mais entre a chegada dos veículos?

36. A durabilidade (em horas) de um dispositivo eletrônico é uma variável aleatória com a seguinte função densidade exponencial de probabilidade:

$$f(x) = \frac{1}{50} e^{-x/50} \quad \text{para } x \geq 0$$

a. Qual é durabilidade média do dispositivo?
b. Qual é a probabilidade de o dispositivo falhar nas primeiras 25 horas de operação?
c. Qual é a probabilidade de o dispositivo operar 100 horas ou mais antes de falhar?

37. A Sparagowski & Associates realizou um estudo dos tempos de atendimento nos guichês de lanchonetes com serviços de *drive-thru*. O tempo médio entre a colocação de um pedido e o seu recebimento no McDonald's foi de três minutos e 18 segundos (*The Cincinnati Enquirer*, 9 de julho de 2000). Filas de espera como estas freqüentemente seguem uma distribuição exponencial de probabilidade.

a. Qual é a probabilidade de o tempo de atendimento a um cliente ser inferior a 2 minutos?
b. Qual é a probabilidade de o tempo de atendimento a um cliente ser superior a 5 minutos?
c. Qual é a probabilidade de o tempo de atendimento a um cliente ser superior a 3 minutos e 18 segundos?

38. De acordo com uma pesquisa intitulada Primary Reader Survey, promovida pela revista *Barron's*, 30 é o número médio anual de transações de investimentos feitas por um assinante (www.barronsmag.com, 28 de julho de 2000). Suponha que o número de transações em um ano siga a distribuição de probabilidade de Poisson.

a. Apresente a distribuição de probabilidade correspondente ao intervalo de tempo entre as transações de investimento.
b. Qual é a probabilidade de não ocorrer nenhuma transação durante o mês de janeiro em relação a um assinante em particular?
c. Qual é a probabilidade de a próxima transação ocorrer dentro da próxima quinzena em relação a um assinante em particular?

Resumo

Este capítulo ampliou a discussão das distribuições de probabilidade para o caso das variáveis aleatórias contínuas. A principal diferença conceitual entre as distribuições discretas e as distribuições de probabilidade contínuas envolve o método de se calcular probabilidades. No que refere às distribuições discretas, a função de probabilidade $f(x)$ fornece a probabilidade de a variável aleatória x assumir valores diversos. Quanto às distribuições contínuas, a função densidade de probabilidade $f(x)$ não produz valores probabilísticos diretamente. Ao contrário, as probabilidades são fornecidas pelas áreas sob a curva ou gráfico da função densidade de probabilidade $f(x)$. Uma vez que a área sob a curva acima de um ponto simples é zero, observamos que a probabilidade de qualquer valor em particular também é zero, quando se trata de uma variável aleatória contínua.

Três distribuições contínuas de probabilidade foram tratadas detalhadamente: a distribuição uniforme, a distribuição normal e a distribuição exponencial. A distribuição normal é amplamente empregada na inferência estatística e será extensamente usada no restante deste livro.

Glossário

Função densidade de probabilidade Uma função usada para calcular as probabilidades de uma variável aleatória contínua. A área sob o gráfico de uma função densidade de probabilidade ao longo de um intervalo representa a probabilidade.

Distribuição uniforme de probabilidade Uma distribuição contínua de probabilidade em que a probabilidade de a variável aleatória assumir um valor em um intervalo qualquer é a mesma para cada intervalo de igual extensão.

Distribuição normal de probabilidade Uma distribuição contínua de probabilidade. Sua função densidade de probabilidade tem a forma de sino e é determinada por sua média μ e pelo desvio padrão σ.

Distribuição normal padrão de probabilidade Uma distribuição normal com média 0 (zero) e desvio padrão 1.

Fator de correção de continuidade O valor 0,5 que é adicionado e/ou subtraído de um valor de x quando a distribuição normal contínua de probabilidade é utilizada para fazer a aproximação à distribuição binomial discreta.

Distribuição exponencial de probabilidade Uma distribuição contínua de probabilidade que é útil para calcular probabilidades referentes ao tempo necessário para se concluir uma tarefa.

Fórmulas-Chave

Função Densidade Uniforme de Probabilidade

$$f(x) = \begin{cases} \dfrac{1}{b-a} & \text{para } a \leq x \leq b \\ 0 & \text{outro ponto qualquer} \end{cases} \qquad (6.1)$$

Função de Densidade Normal de Probabilidade

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2} \qquad (6.2)$$

Como Converter em Distribuição Normal Padrão

$$z = \frac{x-\mu}{\sigma} \qquad (6.3)$$

Função Densidade Exponencial de Probabilidade

$$f(x) = \frac{1}{\mu} e^{-x/\mu} \quad \text{para } x \geq 0, \mu > 0 \qquad (6.4)$$

Distribuição Exponencial: Probabilidades Cumulativas

$$P(x \leq x_0) = 1 - e^{-x_0/\mu} \qquad (6.5)$$

Exercícios Suplementares

39. Uma executiva de negócios, transferida de Chicago para Atlanta, precisa vender rapidamente sua casa em Chicago. O empregador da executiva ofereceu-se para comprar a casa por US$ 210 mil, mas a oferta se encerra no fim da semana. A executiva não tem atualmente uma oferta melhor, mas tem condições de deixar a casa à venda por mais um mês. Em virtude das conversas que manteve com seu corretor de imóveis, a executiva acredita que o preço que obterá se deixar a casa à venda por mais um mês se distribui uniformemente entre US$ 200 mil e US$ 225 mil.
    a. Se ela deixar a casa à venda por mais um mês, qual é a expressão matemática correspondente à função densidade de probabilidade do preço de venda?
    b. Se ela deixar a casa à venda por mais um mês, qual é a probabilidade de vir a obter, no mínimo, US$ 215 mil pela casa?
    c. Se ela deixar a casa à venda por mais um mês, qual é a probabilidade de vir a obter menos de US$ 210 mil?
    d. A executiva deve deixar a casa à venda por mais um mês? Por quê?
40. O U.S. Bureau of Labor Statistics relata que o dispêndio anual médio em alimentação e bebidas referente a todas as famílias é US$ 5.700 (*Money*, dezembro de 2003). Suponha que os gastos anuais com alimentação e bebidas tenham uma distribuição normal e que o desvio padrão seja US$ 1.500.
    a. Quanto 10% das famílias que têm o menor nível de gastos despendem anualmente em alimentação e bebidas?
    b. Qual porcentagem de famílias gasta anualmente mais de US$ 7 mil com alimentação e bebidas?
    c. Quanto 5% das famílias que têm o maior nível de gastos despendem anualmente em alimentação e bebidas?
41. A Motorola usou a distribuição normal para determinar a probabilidade de defeitos e o número de defeitos esperados em um processo de produção. Suponha que um processo de produção produza itens com um peso médio de 10 onças (283,49 g). Calcule a probabilidade de ocorrer um defeito e o número esperado de defeitos em uma rodada de produção de mil unidades, nas seguintes situações:
    a. O desvio padrão do processo é 0,15 e o controle do processo foi ajustado em mais ou menos um desvio padrão. As unidades com pesos inferiores a 9,85 onças (279,24 g) ou superiores a 10,15 onças (287,74 g) serão classificadas como defeitos.
    b. Através de melhorias no projeto dos processos, o desvio padrão do processo pode ser reduzido para 0,05. Suponha que o controle do processo permaneça o mesmo, com os pesos inferiores a 9,85 onças (279,24 g) ou superiores a 10,15 onças (287,74 g) sendo considerados defeitos.
    c. Qual é a vantagem de se reduzir a variação no processo e ajustar, portanto, os limites do controle de processo em um número maior de desvios padrão da média?
42. A quantia anual média que as famílias norte-americanas gastam com o transporte diário é US$ 6.312 (*Money*, agosto de 2001). Suponha que a quantia média tenha uma distribuição normal.
    a. Suponha que você saiba que 4% das famílias norte-americanas gastam menos de US$ 1.000 com o transporte diário. Qual é o desvio padrão da quantia gasta?
    b. Qual é a probabilidade de uma família gastar entre US$ 4 mil e US$ 6 mil?
    c. Qual é a quantia gasta por 3% das famílias que têm os custos de transporte diário mais elevados?
43. A *Condé Nast Traveler* publica uma Gold List dos melhores hotéis de todos os lugares do mundo. O Broadmoor Hotel, em Colorado Springs, possui 700 quartos e está na Gold List de 2004 (*Condé Nast Traveler*, janeiro de 2004). Suponha que a equipe de marketing do Broadmoor preveja uma demanda por 670 quartos para o próximo fim de semana. Suponha que a demanda para o próximo fim de semana tenha uma distribuição normal com um desvio padrão igual a 30.
    a. Qual é a probabilidade de todos os quartos do hotel serem alugados?
    b. Qual é a probabilidade de 50 ou mais quartos não serem alugados?
    c. Você recomendaria que a direção do hotel oferecesse uma promoção para aumentar a procura? Quais considerações seriam importantes?
44. A Ward Doering Auto Sales está pensando em oferecer um contrato de serviço especial que cubra o custo total de qualquer trabalho de reparo necessário nos veículos alugados. Por experiência, o gerente da empresa estima que os custos de reparo anuais estão distribuídos de maneira aproximadamente normal, com uma média de US$ 150 e desvio padrão de US$ 25.

a. Se a empresa oferecer o contrato de serviço aos clientes por um custo anual de US$ 200, qual é a probabilidade de os custos de serviço a qualquer cliente em particular ultrapassarem o preço de contrato de US$ 200?
b. Qual é o lucro esperado da Ward por contrato de serviço?

45. Deixar de dormir o suficiente está causando mortes no trânsito? Um estudo realizado sob o patrocínio da National Highway Traffic Safety Administration descobriu que o número médio de acidentes fatais provocados por motoristas sonolentos anualmente era 1.550 (*Business Week*, 26 de janeiro de 2004). Suponha que o número anual de acidentes fatais por ano esteja distribuído normalmente, com um desvio padrão de 300.
a. Qual é a probabilidade de haver menos de mil acidentes fatais em um ano?
b. Qual é a probabilidade de o número de acidentes fatais situar-se entre mil e 2 mil por ano?
c. Para que um ano se situe entre os 5% máximos com respeito ao número de acidentes fatais, quantos acidentes desse tipo teriam de ocorrer?

46. Considere que as pontuações obtidas nos exames de admissão à universidade estejam normalmente distribuídas, sendo a média 450 e o desvio padrão 100.
a. Qual porcentagem das pessoas que fizeram os exames obtiveram pontuações entre 400 e 500?
b. Suponha que alguém receba a pontuação 630. Das pessoas que fizeram os exames, qual porcentagem obteve uma pontuação melhor? Qual porcentagem obteve uma pontuação pior?
c. Se uma universidade em particular não admitir ninguém com pontuações abaixo de 480, qual porcentagem das pessoas que fizeram os exames seriam aceitas nessa universidade?

47. De acordo com a *Advertising Age*, o salário-base médio das mulheres que trabalham como *copywriters*[4] em firmas de publicidade é mais alto que o salário-base médio dos homens. O salário-base médio das mulheres é US$ 67 mil e o salário-base médio dos homens é US$ 65 mil (*Working Woman*, julho/agosto de 2000). Considere que os salários estão distribuídos normalmente e que o desvio padrão é US$ 7 mil tanto para os homens como para as mulheres.
a. Qual é a probabilidade de uma mulher receber um salário acima de US$ 75 mil?
b. Qual é a probabilidade de um homem receber um salário acima de US$ 75 mil?
c. Qual é a probabilidade de uma mulher receber um salário abaixo de US$ 50 mil?
d. Quanto uma mulher teria de ganhar para ter um salário mais alto que 99% de suas contrapartes do sexo masculino?

48. Uma máquina preenche recipientes com determinado produto. Por experiência, sabe-se que o desvio padrão dos volumes de preenchimento é 0,6 onças (17,74 ml). Se somente 2% dos recipientes contêm menos de 18 onças (532,32 ml), qual é o volume médio de preenchimento efetuado pela máquina? Ou seja, qual deve ser o valor de μ? Considere que os volumes de preenchimento apresentam uma distribuição normal.

49. Considere um exame de múltipla escolha com 50 questões. Cada questão tem quatro respostas possíveis. Suponha que o estudante que tenha feito seu trabalho de casa e participado de todas as aulas tenha 0,75 de probabilidade de responder corretamente a qualquer questão.
a. Um estudante deve responder corretamente a 43 questões ou mais para obter uma nota A. Qual porcentagem dos estudantes que fizeram seus trabalhos de casa e participaram das aulas obterá notas A neste exame de múltipla escolha?
b. O estudante que responder corretamente a um número de 35 a 39 questões receberá uma nota C. Qual porcentagem dos estudantes que fizeram seus trabalhos de casa e participaram das aulas obterá notas C neste exame de múltipla escolha?
c. Um estudante deve responder corretamente a 30 questões ou mais para ser aprovado no exame. Qual porcentagem dos estudantes que fizeram seus trabalhos de casa e participaram das aulas será aprovada no exame?
d. Considere que um estudante não tenha participado das aulas e não tenha feito o trabalho de casa exigido pelo curso. Além disso, suponha que o estudante simplesmente "chutou" as respostas a cada questão. Qual é a probabilidade de esse estudante responder corretamente a 30 questões ou mais e ser aprovado no exame?

[4] NT: *Copywriter*: Redator(a) de texto para anúncios ou matéria promocional.

50. Um jogador de *blackjack*[5] em um cassino de Las Vegas soube que a casa oferecerá um quarto gratuito se o jogo se estender a quatro horas com uma aposta média de US$ 50. A estratégia do jogador apresenta uma probabilidade de 0,49 de ele ganhar qualquer "mão" do jogo, e o jogador sabe que 60 "mãos" são jogadas por hora. Suponha que ele jogue durante quatro horas, com apostas de US$ 50 por mão.
   a. Qual é a expectativa de ganho do jogador?
   b. Qual é a probabilidade de o jogador perder US$ 1.000 ou mais?
   c. Qual é a probabilidade de o jogador ganhar?
   d. Suponha que o jogador inicie com US$ 1.500. Qual é a probabilidade de ele perder todo o dinheiro?
51. O tempo em minutos durante o qual um estudante usa um terminal de computador no centro de informática de uma grande universidade segue uma distribuição exponencial de probabilidade com uma média de 36 minutos. Suponha que um estudante chegue ao terminal exatamente no momento em que outro estudante começa a trabalhar nele.
   a. Qual é a probabilidade de o tempo de espera do segundo estudante ser de 15 minutos ou menos?
   b. Qual é a probabilidade de o tempo de espera do segundo estudante se situar entre 15 e 45 minutos?
   c. Qual é a probabilidade de o segundo estudante ter de esperar uma hora ou mais?
52. O *website* da empresa Bed and Breakfast Inns of North America (www.bestinns.net) tem aproximadamente sete visitantes por minuto (*Time*, setembro de 2001). Suponha que o número de visitas por minuto ao site siga uma distribuição de probabilidade de Poisson.
   a. Qual é o tempo médio entre as visitas ao site?
   b. Apresente a função densidade de probabilidade referente ao tempo entre as visitas ao site.
   c. Qual é a probabilidade de ninguém acessar o site no período de 1 minuto?
   d. Qual é a probabilidade de ninguém acessar o site no período de 12 segundos?
53. O tempo médio de viagem que os residentes na cidade de Nova York gastam para ir ao trabalho é 36,5 minutos (*Time Almanac*, 2001).
   a. Suponha que a distribuição exponencial de probabilidade seja aplicável e apresente a função densidade de probabilidade correspondente ao tempo de viagem que um nova-iorquino típico gasta para ir ao trabalho.
   b. Qual é a probabilidade de um nova-iorquino típico gastar entre 20 e 40 minutos para ir ao trabalho?
   c. Qual é a probabilidade de um nova-iorquino típico gastar mais de 40 minutos para ir ao trabalho?
54. O tempo decorrido (em minutos) entre as chamadas telefônicas em um escritório de reclamações de seguro freqüentemente tem a seguinte distribuição exponencial de probabilidade:

$$f(x) = 0{,}50e^{-0{,}50x} \quad \text{para} \quad x \geqslant 0$$

   a. Qual é o tempo médio entre as chamadas telefônicas?
   b. Qual é a probabilidade de haver 30 segundos ou menos entre as chamadas telefônicas?
   c. Qual é a probabilidade de haver 1 minuto ou menos entre as chamadas telefônicas?
   d. Qual é a probabilidade de haver 5 minutos ou mais sem chamadas telefônicas?

Estudo de Caso – Specialty Toys

A Specialty Toys, Inc., vende uma grande variedade de novos e inovadores brinquedos infantis. A gerência percebeu que a temporada que antecede as festas de fim de ano é a melhor época para lançar um novo brinquedo no mercado, uma vez que é nesse período que muitas famílias procuram novas idéias de presentes para as comemorações de dezembro. Quando a Specialty descobre um novo brinquedo com bom potencial de mercado, escolhe uma data em outubro para efetuar a entrada no mercado.

Para colocar os brinquedos em suas lojas até outubro, a Specialty faz os seus pedidos aos fabricantes de uma só vez no mês de junho ou julho de cada ano. A demanda por brinquedos infantis pode ser altamente volátil. Se um novo brinquedo obtiver grande sucesso, a sensação de escassez no mercado freqüentemente aumenta a demanda a níveis elevados, e grandes lucros podem ser percebidos. Entretanto, novos

[5] NT: *Blackjack* – O *blackjack*, ou "vinte-e-um", é um jogo de azar muito popular nos cassinos de Las Vegas.

brinquedos também podem encalhar, deixando a Specialty entulhada de grandes níveis de estoque que precisam ser vendidos a preços reduzidos. A questão mais importante que a empresa enfrenta é decidir quantas unidades de um novo brinquedo devem ser adquiridas para satisfazer à demanda de vendas prevista. Se comprar muito pouco, perderá vendas; se comprar demais, os lucros serão reduzidos em razão dos baixos preços realizados nas vendas para limpar o estoque.

Para a próxima temporada, a Specialty planeja lançar no mercado um novo produto, chamado Weather Teddy. Essa variação de ursinho falante é produzida por uma empresa de Taiwan. Quando a criança pressiona a mão do ursinho, ele começa a falar. Um barômetro embutido seleciona uma das cinco respostas que dão uma previsão do tempo. As respostas variam de "Parece que o dia está muito bonito! Divirta-se!" a "Acho que pode chover hoje. Não se esqueça do seu guarda-chuva!". Os testes realizados com o produto mostram que, embora não seja uma previsão meteorológica perfeita, suas previsões do tempo são surpreendentemente boas. Diversos gerentes da Specialty afirmaram que o Teddy faz previsões do tempo tão boas quanto muitas das previsões meteorológicas locais apresentadas na televisão.

À semelhança do que ocorre com outros produtos, a Specialty se defronta com a decisão de quantas unidades de Weather Teddy encomendar para o próximo período de festas. Membros da equipe administrativa sugeriram encomendar quantidades de 15 mil, 18 mil, 24 mil ou 28 mil unidades. A larga margem de lotes de compra sugeridos indica uma considerável discordância em relação ao potencial de mercado. A equipe de gerência de produto pede-lhe uma análise das probabilidades de quebra de estoque (*stock-out*) para os vários lotes de compra, uma estimativa do lucro potencial, e pede-lhe também para auxiliá-la a elaborar uma recomendação de lote de compra. A Specialty espera vender o Weather Teddy por US$ 24, baseando-se em um custo de US$ 16 por unidade. Se houver saldos de estoque depois do período de festas de fim de ano, a Specialty venderá todo o estoque restante a US$ 5 por unidade. Depois de revisar o histórico de vendas de produtos similares, o planejador sênior de vendas previu uma demanda esperada de 20 mil unidades, com 0,90 de probabilidade de a demanda se situar entre 10 mil e 30 mil unidades.

Relatório Administrativo

Prepare um relatório administrativo que encaminhe as seguintes questões e recomende um lote de compra relativo ao produto Weather Teddy.

1. Use a previsão do planejador de vendas para descrever uma distribuição normal de probabilidade que possa ser usada para fazer a aproximação à distribuição da demanda. Faça um esboço da distribuição e apresente a média e o desvio padrão.
2. Calcule a probabilidade de quebra de estoque para os lotes de compra sugeridos pelos membros da equipe administrativa.
3. Calcule o lucro projetado para os lotes de compra sugeridos pela equipe administrativa considerando três cenários: o pior caso, no qual as vendas são de 10 mil unidades, o caso mais provável, em que as vendas são de 20 mil unidades, e o melhor caso, em que as vendas são de 30 mil unidades.
4. Um dos gerentes da Specialty achava que o potencial de lucro era tão grande que o lote de compra poderia ter 70% de chances de satisfazer a demanda e somente 30% de chances de haver uma quebra de estoques. Qual lote deveria ser encomendado sob essa política, e qual é o lucro projetado sob os três cenários de vendas?
5. Apresente sua própria recomendação de lote de compra e anote as projeções de lucro associadas. Forneça um fundamento lógico para sua recomendação.

Apêndice 6.1 – Distribuições Contínuas de Probabilidade com o Minitab

Vamos demonstrar o procedimento para se calcular probabilidades contínuas com o Minitab reportando-nos ao problema da Grear Tire Company, em que a durabilidade dos pneus em termos de milhas foi descrita por uma distribuição normal, com $\mu = 36.500$ e $\sigma = 5.000$. Uma das questões foi: qual é a probabilidade de a durabilidade dos pneus em milhas ultrapassar 40 mil milhas (64.373 quilômetros)?

Em relação às distribuições contínuas de probabilidade, o Minitab fornece uma probabilidade cumulativa; isto é, o Minitab oferece a probabilidade de a variável aleatória assumir um valor menor ou igual a

uma constante específica. Quanto à questão da durabilidade dos pneus da Grear, o Minitab pode ser usado para determinar a probabilidade cumulativa de a durabilidade em milhas ser menor ou igual a 40 mil milhas (a constante específica, nesse caso, 40 mil). Depois de obtermos a probabilidade cumulativa do Minitab, precisamos subtraí-la de 1 para determinar a probabilidade de a durabilidade do pneu ultrapassar 40 mil milhas.

Antes de usar o Minitab para calcular uma probabilidade, precisamos inserir a constante específica em uma coluna da planilha. Quanto à questão da durabilidade dos pneus da Grear, inserimos a constante específica 40 mil na coluna C1 da planilha do Minitab. As etapas para usar o Minitab para calcular a probabilidade cumulativa da variável aleatória normal, considerando um valor menor ou igual a 40 mil, são apresentadas a seguir:

Etapa 1. Selecione o menu **Calc**
Etapa 2. Escolha **Probability Distributions**
Etapa 3. Escolha a opção **Normal**
Etapa 4. Quando a caixa de diálogo Distribuição Normal aparecer:
Selecione **Cumulative probability**
Digite 36.500 na caixa **Mean**
Digite 5.000 na caixa **Standard deviation**
Digite C1 na caixa **Input column** (a coluna que contém 40.000)
Dê um clique em **OK**

Depois que o usuário dá um clique em **OK**, o Minitab imprime a probabilidade cumulativa de a variável aleatória normal assumir um valor menor ou igual a 40 mil. O Minitab mostra que essa probabilidade é de 0,7580. Como estamos interessados na probabilidade de a durabilidade do pneu ser maior que 40 mil, a probabilidade desejada é 1 – 0,7580 = 0,2420.

Uma segunda questão no problema da Grear Tire Company foi: qual garantia de durabilidade em milhas a Grear deve fixar para assegurar que não mais de 10% dos pneus se qualifiquem à garantia? Aqui nos é dada uma probabilidade e queremos descobrir o valor correspondente da variável aleatória. O Minitab usa uma rotina de cálculo inversa para encontrar o valor da variável aleatória associada a determinada probabilidade cumulativa. Primeiramente, precisamos introduzir a probabilidade cumulativa em uma coluna da planilha do Minitab (digamos, C1).

Nesse caso, a probabilidade cumulativa desejada é 0,10. Depois, as três primeiras etapas de procedimento do Minitab são idênticas às que já foram relatadas. Na etapa 4, selecionamos **Inverse cumulative probability** em vez de **Cumulative probability** e concluímos as partes restantes da etapa. O Minitab exibirá, então, a garantia de durabilidade de 30.092 milhas (48.428 km).

O Minitab é capaz de calcular probabilidades para outras distribuições contínuas de probabilidade, até mesmo a distribuição exponencial de probabilidade. Para calcular as probabilidades exponenciais, siga o procedimento apresentado anteriormente referente à distribuição normal de probabilidade e escolha a opção **Exponential** na etapa 3. A etapa 4 é idêntica ao que foi descrito, com exceção de que não é necessário introduzir um desvio padrão. Os dados de saída (*output*) das probabilidades cumulativas e probabilidades cumulativas inversas são idênticos aos da distribuição normal de probabilidade.

Apêndice 6.2 – Distribuições Contínuas de Probabilidade com o Excel

O Excel tem a capacidade de calcular probabilidades de diversas distribuições contínuas de probabilidade, até mesmo as distribuições normal e exponencial. Neste apêndice, descreveremos como o Excel pode ser usado para calcular probabilidades de qualquer distribuição normal de probabilidade. Os procedimentos referentes às distribuições exponenciais e outras distribuições contínuas são idênticos aos que descrevemos com relação à distribuição normal de probabilidade.

Retornemos ao problema da Grear Tire Company, em que a durabilidade dos pneus em termos de milhas foi descrita como uma distribuição normal de probabilidade, com μ = 36.500 e σ = 5.000. Suponha que estejamos interessados na probabilidade de a durabilidade do pneu ultrapassar 40 mil milhas (64.373 km).

A função DIST.NORM do Excel fornece as probabilidades cumulativas de uma distribuição normal. A forma geral da função DIST.NORM (x, μ, σ, cumulativo). Quanto ao quarto argumento, VERDADEIRO é especificado se uma probabilidade cumulativa for desejada. Desse modo, para calcular a probabilidade cumulativa de a durabilidade do pneu ser menor ou igual a 40 mil milhas (64.373 km), introduziríamos a seguinte fórmula em qualquer célula de uma planilha do Excel:

=DIST.NORM(40000;36500;5000;VERDADEIRO)

Neste ponto, aparecerá 0,7580 na célula em que a fórmula foi inserida, indicando que a probabilidade de a durabilidade do pneu ser menor ou igual a 40 mil milhas é 0,7580. Portanto, a probabilidade de a durabilidade do pneu ultrapassar 40 mil milhas é 1 – 0,7580 = 0,2420.

A função DIST.NORM do Excel usa um cálculo inverso para encontrar o valor de x correspondente a determinada probabilidade cumulativa. Por exemplo, suponha que queiramos descobrir qual é o número de milhas que a Grear deve oferecer como garantia a fim de que não mais de 10% dos pneus se qualifiquem à garantia. Digitaríamos a seguinte fórmula em qualquer célula de uma planilha do Excel

=INV.NORM(0,1;35600;5000)

Neste ponto, aparecerá 30.092 na célula em que a fórmula foi inserida, indicando que a probabilidade de um pneu durar 30.092 milhas (48.428 km) é inferior a 0,10.

A função do Excel para calcular probabilidades exponenciais é DISTEXPON. Ela é fácil de usar. Mas se alguém precisar de ajuda para especificar os valores apropriados para os argumentos, a função Inserir do Excel pode ser usada (veja o Apêndice 2.2).

Amostragens e Distribuições Amostrais

ESTATÍSTICA NA PRÁTICA

MEADWESTVACO CORPORATION*
Stamford, Connecticut

A MeadWestvaco Corporation, uma empresa que ocupa a liderança na produção de embalagens, papéis especiais, *coated paper*[1], produtos de consumo, de escritório e de química fina, emprega mais de 30 mil pessoas. Opera internacionalmente em 33 países e atende a clientes localizados em aproximadamente 100 países. A MeadWestvaco detém uma posição de liderança na produção de papéis, com uma capacidade anual de 1,8 milhão de toneladas. Os produtos da empresa incluem papel para livros didáticos, papel brilhante (*glossy*) para revistas, sistemas de embalagem de bebidas e produtos de escritório. A equipe de consultoria interna da MeadWestvaco recorre a amostragens para produzir uma variedade de informações que possibilitam à empresa obter significativos benefícios de produtividade e permanecer competitiva.

Por exemplo, a MeadWestvaco mantém grandes propriedades florestais cujas árvores são a matéria-prima para muitos dos produtos da empresa. Os gerentes necessitam de informações confiáveis e precisas a respeito das áreas de cultivo de madeira e florestas para avaliar a capacidade da empresa para suprir suas necessidades futuras de matéria-prima. Qual é o atual volume das florestas? Qual foi o crescimento das florestas no passado? Qual é a projeção de crescimento futuro das florestas? Com as respostas a essas importantes ques-

* Os autores agradecem ao Dr. Edward P. Winkofsky por fornecer esta "Estatística na Prática".
[1] NT: *Coated paper* – Papel cuja superfície foi tratada para receber impressões em escala de cinza ou colorida.

tões, os gerentes da MeadWestvaco podem desenvolver planos para o futuro, incluindo o planejamento do plantio a longo prazo e a programação de corte das árvores.

Como a MeadWestvaco obtém as informações de que necessita a respeito de suas vastas propriedades florestais? Dados coletados de pequenos lotes amostrais de todas as florestas constituem a base para a empresa tomar conhecimento do número total de árvores que ela possui. Para identificar os lotes amostrais, as áreas florestais são primeiramente divididas em três seções baseadas na localização e nos tipos de árvore. Usando mapas e números aleatórios, os analistas da MeadWestvaco identificam amostras aleatórias de lotes de 1/5 a 1/7 de acre[2] em cada seção da floresta. Os engenheiros florestais da MeadWestvaco coletam dados desses lotes amostrais para conhecer a população de árvores da floresta.

Os engenheiros florestais de toda a organização participam do processo de coleta de dados em campo. Periodicamente, equipes de duas pessoas reúnem as informações obtidas sobre cada árvore de cada um dos lotes amostrais. Os dados amostrais são inseridos no sistema computadorizado denominado Continuous Forest Inventory (CFI) – inventário contínuo de florestas – da empresa. Os relatórios do sistema CFI incluem uma série de sumários de distribuição de freqüência que contêm estatísticas sobre os tipos de árvore, volume atual, taxas de crescimento florestal passadas e projeções do crescimento e volume florestal futuros. A amostragem e os sumários estatísticos dos dados amostrais correspondentes produzem os relatórios que são fundamentais à administração eficaz das florestas e áreas de cultivo de madeira da MeadWestvaco.

Neste capítulo, você aprenderá a amostragem aleatória simples e o processo de escolha da amostra. Além disso, aprenderá como são usados certos métodos estatísticos, como a média amostral e a proporção da amostra, para estimar a média e a proporção da população. Também é introduzido o importante conceito de distribuição amostral.

No Capítulo 1, definimos o que é uma *população* e uma *amostra*. As definições são reapresentadas a seguir:

1. Uma *população* é o conjunto de todos os elementos de interesse em um estudo.
2. Uma *amostra* é um subconjunto da população.

Características numéricas de uma população, por exemplo, a média e o desvio padrão, são chamadas **parâmetros**. Um dos propósitos fundamentais da inferência estatística é desenvolver estimativas e testar hipóteses a respeito dos parâmetros populacionais usando a informação contida em uma amostra.

Vamos iniciar referindo-nos a duas situações nas quais amostras produzem estimativas dos parâmetros populacionais:

1. Um fabricante de pneus desenvolveu um novo tipo de pneu, projetado para proporcionar um aumento da durabilidade em termos de milhas em relação à atual linha de pneus da empresa. Para estimar o número médio de milhas proporcionadas pelos novos pneus, o fabricante selecionou uma amostra de 120 pneus novos para teste. Os resultados do teste produziram uma média amostral de 36.500 milhas (58.741 km). Portanto, uma estimativa do número médio de milhas para a população de novos pneus era de 36.500 milhas.
2. Os membros de um partido político consideravam a possibilidade de apoiar determinado candidato nas eleições ao Senado dos Estados Unidos, e os líderes do partido queriam uma estimativa da proporção de eleitores inscritos favoráveis ao candidato. O tempo e o custo associados ao trabalho de contatar cada indivíduo da população de eleitores inscritos eram proibitivos. Portanto, foi selecionada uma amostra de 400 eleitores inscritos, dos quais 160 indicaram preferência pelo candidato. A estimativa da proporção da população de eleitores inscritos favoráveis ao candidato foi de 160/400 = 0,40.

Esses dois exemplos que acabamos de apresentar ilustram algumas das razões pelas quais se usam amostras. Observe que, no exemplo da durabilidade dos pneus, a coleta de dados sobre a vida útil do pneu envolve gastar cada pneu testado. Evidentemente, não é viável testar cada pneu da população; uma amostra é a única maneira realística de se obter os dados desejados de durabilidade dos pneus. No exemplo envolvendo as eleições, contatar cada eleitor inscrito da população é teoricamente possível, mas o tempo e o custo desse trabalho são por demais proibitivos; desse modo, é preferível uma amostra dos eleitores inscritos.

[2] NT: 1 acre = 40,47 ares.

É importante entender que os resultados da amostra fornecem somente *estimativas* dos valores das características populacionais. Não esperamos que a média amostral de 36.500 milhas (58.741 km) seja exatamente igual ao número médio de milhas para todos os pneus da população nem esperamos que exatamente 0,40, ou 40%, da população de eleitores inscritos seja favorável ao candidato. A razão é simplesmente esta: a amostra contém somente uma parcela da população. Com métodos de amostragem apropriados, os resultados da amostra produzirão "boas" estimativas dos parâmetros populacionais. Mas, qual é o nosso nível de confiança de que os resultados da amostra serão bons? Felizmente, há procedimentos estatísticos disponíveis para responder a essa questão.

Uma média amostral produz uma estimativa da média da população, e uma proporção amostral fornece uma estimativa da proporção da população. Quando se trata de estimativas como estas, alguns erros de estimação podem ser esperados. Este capítulo apresenta a base para que se possa determinar qual poderia ser a extensão desses erros.

Neste capítulo, mostramos como a amostragem aleatória simples pode ser usada para selecionar uma amostra de uma população. Depois, mostraremos como os dados obtidos de uma amostra aleatória simples podem ser usados para se calcular estimativas da média de uma população, o desvio padrão de uma população e a proporção de uma população. Além disso, introduziremos o importante conceito de distribuição amostral. Conforme mostraremos, o conhecimento da distribuição amostral apropriada é que nos possibilita fazer afirmações sobre quão próximas estão as estimativas amostrais dos parâmetros populacionais correspondentes. A última seção discute algumas alternativas à amostragem aleatória simples que freqüentemente são empregadas na prática.

7.1 PROBLEMA DE AMOSTRAGEM DA ELECTRONICS ASSOCIATES

ARQUIVO DA INTERNET
EAI

O diretor de pessoal da Electronics Associates, Inc. (EAI) foi incumbido da tarefa de desenvolver um perfil dos 2.500 gerentes da empresa. As características a serem identificadas incluem o salário médio anual dos gerentes e a proporção de gerentes que concluíram o programa de treinamento gerencial da empresa.

Usando os 2.500 gerentes como a população para esse estudo, podemos encontrar o salário anual e o *status* do programa de treinamento de cada indivíduo consultando os registros de pessoal da empresa. O arquivo de dados que contém essa informação referente a todos os 2.500 gerentes da população encontra-se no site www.thomsonlearning.com.br/estatapl.htm.

Usando o conjunto de dados da EAI e as fórmulas apresentadas no Capítulo 3, calculamos a média populacional e o desvio padrão correspondentes aos dados de salário anual.

Média populacional: μ = US\$ 51.800

Desvio padrão da população: σ = US\$ 4.000

Freqüentemente, o custo para coletar as informações de uma amostra é substancialmente menor que o custo para coletar informações de uma população, especialmente quando é necessário realizar entrevistas pessoais para coletar essas informações.

Os dados referentes ao *status* no programa de treinamento mostram que 1.500 dos 2.500 gerentes concluíram o programa de treinamento. Se admitirmos que p denota a proporção da população que concluiu o programa de treinamento, verificamos que $p = 1.500/2.500 = 0{,}60$. O salário médio anual da população (μ = US\$ 51.800), o desvio padrão do salário anual da população (σ = US\$ 4 mil) e a proporção da população que concluiu o programa de treinamento ($p = 0{,}60$) são parâmetros da população de gerentes da EAI.

Agora, suponha que as informações necessárias sobre todos os gerentes do EAI não estivessem prontamente disponíveis no banco de dados da empresa. A questão que consideramos agora é como o diretor de pessoal da empresa pode obter estimativas dos parâmetros populacionais usando uma amostra de gerentes em vez de todos os 2.500 gerentes da população. Suponha que seja usada uma amostra de 30 gerentes. Evidentemente, o tempo e o custo para desenvolver um perfil seriam substancialmente menores em relação aos 30 gerentes do que para a população inteira. Se o diretor de pessoal pudesse ter a certeza de que a amostra de 30 gerentes forneceria as informações adequadas a respeito da população de 2.500 gerentes, trabalhar com uma amostra seria preferível a trabalhar com a população inteira. Vamos explorar a possibilidade de usar uma amostra para o estudo realizado pela EAI, considerando primeiramente como podemos identificar uma amostra de 30 gerentes.

7.2 AMOSTRAGEM ALEATÓRIA SIMPLES

Diversos métodos podem ser usados para selecionar uma amostra de uma população; um dos mais comuns é a **amostragem aleatória simples**. A definição de amostra aleatória simples e o processo de seleção de

uma amostra aleatória simples dependem de a população ser *finita* ou *infinita*. Como o problema de amostragem da EAI envolve uma população finita de 2.500 gerentes, consideraremos primeiramente a amostragem de uma população finita.

Amostragem de Populações Finitas

Uma amostra aleatória simples de tamanho n de uma população finita de tamanho N é definida da seguinte maneira:

AMOSTRA ALEATÓRIA SIMPLES (POPULAÇÃO FINITA)

Uma amostra aleatória simples de tamanho n de uma população finita de tamanho N é uma amostra selecionada de tal maneira que cada amostra possível de tamanho n tenha a mesma probabilidade de ser escolhida.

Números aleatórios gerados por computador também podem ser usados para implementar o processo de escolha da amostra aleatória. O Excel oferece uma função para gerar números aleatórios em suas planilhas.

Um procedimento para selecionar uma amostra aleatória simples de uma população finita é escolher os elementos da amostra, um a cada vez, de tal maneira que, a cada etapa, cada um dos elementos restantes da população tenha a mesma probabilidade de ser escolhido. Amostrar n elementos dessa maneira satisfará a definição de amostra aleatória simples de uma população finita.

Os números aleatórios da tabela são apresentados em grupos de cinco para facilitar a leitura.

Para selecionar uma amostra aleatória simples da população finita de gerentes da EAI, primeiramente atribuímos um número a cada gerente. Por exemplo, podemos atribuir os números de 1 a 2.500 aos gerentes, na ordem em que seus nomes aparecem no arquivo de pessoal da EAI. Em seguida, consultamos a lista de números aleatórios da Tabela 7.1. Usando a primeira linha da tabela, cada dígito, 6, 3, 2, ..., é um dígito aleatório que tem igual chance de ocorrer. Uma vez que o maior número da lista da população de gerentes da EAI, 2.500, tem quatro dígitos, selecionaremos números aleatórios na tabela, em conjuntos ou grupos de quatro dígitos. Não obstante podermos iniciar a seleção de números aleatórios em qualquer lugar da tabela e nos deslocarmos sistematicamente na direção que preferirmos, usaremos a primeira linha da Tabela 7.1 e nos deslocaremos da esquerda para a direita. Os sete primeiros números aleatórios de quatro dígitos são

6.327 1.599 8.671 7.445 1.102 1.514 1.807

Como os números da tabela são aleatórios, esses números de quatro dígitos são igualmente prováveis.

Tabela 7.1 Números aleatórios

63271	59986	71744	51102	15141	80714	58683	93108	13554	79945
88547	09896	95436	79115	08303	01041	20030	63754	08459	28364
55957	57243	83865	09911	19761	66535	40102	26646	60147	15702
46276	87453	44790	67122	45573	84358	21625	16999	13385	22782
55363	07449	34835	15290	76616	67191	12777	21861	68689	03263
69393	92785	49902	58447	42048	30378	87618	26933	40640	16281
13186	29431	88190	04588	38733	81290	89541	70290	40113	08243
17726	28652	56836	78351	47327	18518	92222	55201	27340	10493
36520	64465	05550	30157	82242	29520	69753	72602	23756	54935
81628	36100	39254	56835	37636	02421	98063	89641	64953	99337
84649	48968	75215	75498	49539	74240	03466	49292	36401	45525
63291	11618	12613	75055	43915	26488	41116	64531	56827	30825
70502	53225	03655	05915	37140	57051	48393	91322	25653	06543
06426	24771	59935	49801	11082	66762	94477	02494	88215	27191
20711	55609	29430	70165	45406	78484	31639	52009	18873	96927
41990	70538	77191	25860	55204	73417	83920	69468	74972	38712
72452	36618	76298	26678	89334	33938	95567	29380	75906	91807
37042	40318	57099	10528	09925	89773	41335	96244	29002	46453
53766	52875	15987	46962	67342	77592	57651	95508	80033	69828
90585	58955	53122	16025	84299	53310	67380	84249	25348	04332
32001	96293	37203	64516	51530	37069	40261	61374	05815	06714
62606	64324	46354	72157	67248	20135	49804	09226	64419	29457
10078	28073	85389	50324	14500	15562	64165	06125	71353	77669
91561	46145	24177	15294	10061	98124	75732	00815	83452	97355
13091	98112	53959	79607	52244	63303	10413	63839	74762	50289

Agora, podemos usar esses números aleatórios de quatro dígitos para dar a cada gerente da população uma chance igual de ser incluído na amostra aleatória. O primeiro número, 6.327, é maior que 2.500. Ele não corresponde a um dos gerentes enumerados da população e, portanto, é descartado. O segundo número, 1.599, está entre 1 e 2.500. Assim, o primeiro gerente selecionado para a amostra aleatória é o número 1.599 na lista dos gerentes da EAI. Continuando o processo, ignoramos os números 8.671 e 7.445 antes de identificar os gerente de número 1.102, 1.514 e 1.807 para serem incluídos na amostra aleatória. Esse processo continua até que a amostra aleatória simples de 30 gerentes da EAI tenha sido obtida.

Ao implementarmos esse processo de seleção da amostra aleatória simples, é possível que um número aleatório usado anteriormente possa aparecer novamente na tabela antes de a amostra de 30 gerentes da EAI ser escolhida. Como não queremos selecionar um mesmo gerente mais de uma vez, quaisquer números aleatórios utilizados anteriormente são ignorados, pois o gerente correspondente já foi incluído na amostra. O ato de selecionarmos uma amostra dessa maneira é chamado **amostragem sem substituição**. Se selecionarmos uma amostra de tal maneira que sejam aceitáveis números aleatórios escolhidos anteriormente e gerentes específicos possam ser incluídos na amostra duas ou mais vezes, estaremos realizando uma **amostragem com substituição**. A amostragem com substituição é uma maneira válida de se identificar uma amostra aleatória simples. Entretanto, a amostragem sem substituição é o procedimento de amostragem usado com maior freqüência. Quando nos referirmos à amostragem aleatória simples, o pressuposto é de que se trata de uma amostragem sem substituição.

Na prática, uma população a ser estudada geralmente é considerada infinita quando envolve um processo contínuo que impossibilita a listagem ou a contagem de cada elemento da população.

Amostragem de Populações Infinitas

Em algumas situações, a população ou é infinita ou tão grande que, para fins práticos, precisa ser tratada como infinita. Por exemplo, suponha que um restaurante de *fast-food* queira obter um perfil dos seus clientes selecionando uma amostra aleatória simples de clientes e pedindo a cada um para preencher um breve questionário.

Nesse tipo de situação, o processo contínuo de visitas de clientes ao restaurante pode ser visto como proveniente de uma população contínua. A definição de amostra aleatória simples de uma população infinita é a seguinte:

AMOSTRA ALEATÓRIA SIMPLES (POPULAÇÃO INFINITA)

Uma amostra aleatória simples de uma população infinita é uma amostra selecionada de tal maneira que as condições seguintes sejam satisfeitas:

1. Cada elemento selecionado vem dessa população.
2. Cada elemento é selecionado de maneira independente.

Quanto ao exemplo de selecionar uma amostra aleatória simples de clientes de um restaurante de *fast-food*, a primeira condição é satisfeita por qualquer cliente que entre no restaurante. A segunda condição é satisfeita selecionando-se clientes independentemente. O propósito da segunda condição é impedir que haja um viés na seleção. Ocorreria um viés de seleção se, por exemplo, cinco clientes consecutivos selecionados fossem, todos, amigos entre si que chegassem juntos ao restaurante. Poderíamos esperar que esses clientes apresentassem perfis semelhantes. O viés de seleção pode ser evitado assegurando-se de que a escolha de um cliente em particular não influa na escolha de outro cliente qualquer. Em outras palavras, os clientes devem ser escolhidos de maneira independente.

Quanto às populações infinitas, um procedimento de seleção de amostras deve ser idealizado especialmente para selecionar os itens de maneira independente e, desse modo, evitar um viés de seleção que possa atribuir maiores probabilidades de escolha a certos tipos de elemento.

O McDonald's, líder no ramo de restaurantes de *fast-food*, implementou um procedimento de amostragem aleatória simples exatamente para esse tipo de situação. O procedimento de amostragem se baseou no fato de que alguns clientes apresentavam cupons de desconto. Quando queria que um cliente apresentasse um cupom de descontos, o cliente era servido e, em seguida, solicitado a preencher um questionário de perfil do cliente. Uma vez que os clientes que chegavam apresentavam cupons de desconto aleatoriamente e de maneira independente, esse esquema de amostragem assegurava que os clientes eram selecionados independentemente. Desse modo, as duas condições necessárias a uma amostra aleatória simples de uma população infinita eram satisfeitas.

Populações infinitas freqüentemente estão associadas a processos ininterruptos que operam continuamente ao longo do tempo. Por exemplo, peças que são manufaturadas em uma linha de produção, as transações financeiras que ocorrem em um banco, as chamadas telefônicas a um centro de suporte técnico, e clientes que entram em uma loja, todos, podem ser vistos como integrantes de uma população infinita. Nesses casos, um procedimento criativo de amostragem garantirá que não ocorra nenhum viés de seleção e que os elementos da amostra são selecionados de maneira independente.

NOTAS E COMENTÁRIOS

1. O número de diferentes amostras aleatórias simples de tamanho n que podem ser selecionadas de uma população infinita de tamanho N é

$$\frac{N!}{n!(N-n)!}$$

Nessa fórmula, $N!$ e $n!$ são os cálculos fatoriais discutidos no Capítulo 4. Em relação ao problema da EAI, com $N = 2.500$ e $n = 30$, essa expressão pode ser usada para mostrar que aproximadamente $2{,}74 \times 10^{69}$ diferentes amostras aleatórias simples de 30 gerentes da EAI podem ser obtidas.

2. Softwares de computador podem ser usados para selecionar uma amostra aleatória. Nos apêndices do capítulo, mostramos como o Minitab e o Excel podem ser utilizados para selecionar uma amostra aleatória simples de uma população infinita.

Exercícios

Métodos

AUTOTESTE

1. Considere uma população finita com cinco elementos rotulados A, B, C, D e E. Dez possíveis amostras aleatórias simples de tamanho 2 podem ser selecionadas.
   a. Relacione as dez amostras, iniciando com AB, AC e assim por diante.
   b. Usando a amostragem aleatória simples, qual é a probabilidade de cada amostra de tamanho 2 ser selecionada?
   c. Considere que o número aleatório 1 corresponde a A, o número aleatório 2 corresponde a B e assim por diante. Relacione a amostra aleatória simples de tamanho 2 que será selecionada usando-se os dígitos aleatórios 8 0 5 7 5 3 2.
2. Considere que uma população finita tenha 350 elementos.Usando os três últimos dígitos de cada um dos seguintes números aleatórios de cinco dígitos apresentados a seguir (601, 022, 448, ...), determine os quatro primeiros elementos que serão selecionados para a amostra aleatória simples.

   98.601 73.022 83.448 02.147 34.229 27.553 84.147 93.289 14.209

Aplicações

AUTOTESTE

3. A revista *Fortune* publica dados sobre vendas, lucros, ativos, lucro líquido dos acionistas, valor de mercado e rendimentos por ação das 500 maiores corporações industriais norte-americanas (*Fortune* 500, 2003). Suponha que você queira selecionar uma amostra aleatória simples de dez corporações da lista da *Fortune* 500. Use os três últimos dígitos da coluna 9 da Tabela 7.1, iniciando com 554. Leia a coluna de cima para baixo e identifique os números das dez empresas que seriam selecionadas.
4. Os dez títulos financeiros mais ativos nas Bolsas de Nova York (Nyse), Nasdaq e American (Amex) com capitalizações de mercado acima de US$ 500 milhões são os seguintes (*The Wall Street Journal*, 21 de fevereiro de 2003):

Applied Materials	Nasdaq 100
Cisco Systems	Nextel
Intel	Oracle
Lucent Technologies	SPDR
Microsoft	Sun Microsystems

   a. Suponha que uma amostra aleatória de cinco títulos financeiros sejam selecionados para um estudo detalhado do comportamento dos negócios. Iniciando com o primeiro dígito aleatório da Tabela 7.1 e lendo a coluna de cima para baixo, use os números aleatórios de um único dígito para selecionar uma amostra aleatória simples de cinco títulos financeiros a serem usados nesse estudo.
   b. De acordo com a informação de Notas e Comentários, quantas amostras aleatórias simples de tamanho 5 podem ser selecionadas da lista de dez títulos financeiros?

5. Um grêmio estudantil está interessado em avaliar a proporção de estudantes que são favoráveis à política obrigatória de graduação "*pass-fail*"[3] para cursos eletivos. Uma lista de nomes e endereços dos 645 estudantes matriculados no atual semestre está disponível na secretaria da escola. Usando números aleatórios de três dígitos da linha 10 da Tabela 7.1 e deslocando-se da esquerda para a direita, identifique os dez primeiros estudantes que seriam selecionados usando-se a amostragem aleatória simples. Os números aleatórios de três dígitos iniciam-se com 816, 283 e 610.
6. O *County and City Data Book*, publicado pelo Census Bureau (Departamento do Censo), relaciona informações sobre 3.139 municípios de todo o território norte-americano. Considere que um estudo em nível nacional faça a coleta de dados de 30 municípios escolhidos aleatoriamente. Use números aleatórios de quatro dígitos da última coluna da Tabela 7.1 para identificar os números correspondentes aos cinco primeiros municípios selecionados para a amostra. Ignore os primeiros dígitos e inicie com os números aleatórios de quatro dígitos 9.945, 8.364, 5.702 etc.
7. Suponha que queiramos identificar uma amostra aleatória simples de 12 dos 372 médicos de determinada cidade. Os nomes dos médicos estão disponíveis em uma organização médica local. Use a oitava coluna de números aleatórios de cinco dígitos da Tabela 7.1 para identificar os 12 médicos da amostra. Ignore os dois primeiros dígitos aleatórios de cada agrupamento de cinco dígitos dos números aleatórios. Esse processo inicia-se com o número aleatório 108 e prossegue coluna abaixo na lista de números aleatórios.
8. A relação a seguir apresenta os 25 melhores times de futebol americano da NCAA da temporada de 2002 (*NCAA News*, 4 de janeiro de 2003). Use a nona coluna dos números aleatórios da Tabela 7.1, que se inicia com 13.554, para selecionar uma amostra aleatória simples de seis times de futebol. Comece com o time 13 e use os dois primeiros dígitos de cada linha da nona coluna para realizar o seu processo de seleção. Quais são os seis times de futebol americano selecionados para a amostra aleatória simples?

1. Ohio State	14. Virginia Tech
2. Miami	15. Penn State
3. Georgia	16. Auburn
4. Southern California	17. Notre Dame
5. Oklahoma	18. Pittsburgh
6. Kansas State	19. Marshall
7. Texas	20. West Virginia
8. Iowa	21. Colorado
9. Michigan	22. TCU
10. Washington State	23. Florida State
11. North Carolina State	24. Florida
12. Boise State	25. Virginia
13. Maryland	

9. O *The Wall Street Journal* publica o valor patrimonial líquido, o retorno percentual anual até o presente e o retorno percentual de três anos de 555 fundos mútuos (*The Wall Street Journal*, 25 de abril de 2003). Suponha que uma amostra aleatória simples de 12 dos 555 fundos mútuos seja selecionada para um estudo de acompanhamento do tamanho e desempenho dos fundos mútuos. Use a quarta coluna de números aleatórios da Tabela 7.1, que se inicia em 51.102, para selecionar a amostra aleatória simples de 12 fundos mútuos. Inicie com o fundo mútuo 102 e use os três *últimos* dígitos de cada linha da quarta coluna em seu processo de seleção. Quais são os números dos 12 fundos mútuos da amostra aleatória simples?
10. Indique se as populações a seguir devem ser consideradas finitas ou infinitas:
    a. Todos os eleitores inscritos do estado da Califórnia.
    b. Todos os aparelhos de televisão que poderiam ser produzidos pelo parque industrial da TV-M Company, em Allentown, Pensilvânia.
    c. Todos os pedidos que poderiam ser processados por uma empresa de encomenda postal.
    d. Todas as chamadas telefônicas de emergência que poderiam ser feitas a uma delegacia de polícia local.
    e. Todos os componentes que a Fibercon, Inc., produziu no segundo turno de trabalho no dia 17 de maio.

[3] NT: *Pass-fail*: *Educ.* – Designa um sistema de graduação (notas) no qual um "*pass*" (aprovado) ou um "*fail*" (reprovado) é registrado, em vez de uma nota numérica ou letra (Estados Unidos).

7.3 ESTIMAÇÃO POR PONTO

Agora que descrevemos como selecionar uma amostra aleatória simples, retornemos ao problema da EAI. Uma amostra aleatória simples de 30 gerentes, contendo os dados correspondentes aos salários anuais e à participação no programa de treinamento gerencial, é apresentada na Tabela 7.2. A notação x_1, x_2 etc. é usada para denotar o salário anual do primeiro gerente da amostra, o salário anual do segundo gerente da amostra e assim por diante. A participação no programa de treinamento gerencial é indicada por um Sim na coluna correspondente.

Para estimar o valor do parâmetro de uma população, calculamos uma característica correspondente da amostra, denominada **estatística amostral**. Por exemplo, para estimar a média da população μ e o desvio padrão da população σ referentes aos salários anuais dos gerentes da EAI, usamos os dados da Tabela 7.2 para calcular as estatísticas amostrais correspondentes: a média amostral $\bar{x}$ e o desvio padrão da amostra$s$.

Tabela 7.2 Os salários anuais e a situação no programa de treinamento gerencial referentes a uma amostra aleatória simples de 30 gerentes da EAI

Salário Anual (US$)	Programa de Treinamento Gerencial	Salário Anual (US$)	Programa de Treinamento Gerencial
x_1 = 49.094,30	Sim	x_{16} = 51.766,00	Sim
x_2 = 53.263,90	Sim	x_{17} = 52.541,30	Não
x_3 = 49.643,50	Sim	x_{18} = 44.980,00	Sim
x_4 = 49.894,90	Sim	x_{19} = 51.932,60	Sim
x_5 = 47.621,60	Não	x_{20} = 52.973,00	Sim
x_6 = 55.924,00	Sim	x_{21} = 45.120,90	Sim
x_7 = 49.092,30	Sim	x_{22} = 51.753,00	Sim
x_8 = 51.404,40	Sim	x_{23} = 54.391,80	Não
x_9 = 50.957,70	Sim	x_{24} = 50.164,20	Não
x_{10} = 55.109,70	Sim	x_{25} = 52.973,60	Não
x_{11} = 45.922,60	Sim	x_{26} = 50.241,30	Não
x_{12} = 57.268,40	Não	x_{27} = 52.793,90	Não
x_{13} = 55.688,80	Sim	x_{28} = 50.979,40	Sim
x_{14} = 51.564,70	Não	x_{29} = 55.860,90	Sim
x_{15} = 56.188,20	Não	x_{30} = 57.309,10	Não

Usando as fórmulas da média amostral e do desvio padrão de uma amostra apresentados no Capítulo 3, a média amostral é

$$\bar{x} = \frac{\Sigma x_i}{n} = \frac{1.554.420}{30} = \$\ 51.814$$

e o desvio padrão da amostra é

$$s = \sqrt{\frac{\Sigma(x_i - \bar{x})^2}{n-1}} = \sqrt{\frac{325.009.260}{29}} = \$\ 3.348$$

Para estimar p, que é a proporção de gerentes da população que concluíram o programa de treinamento gerencial, usamos a proporção amostral correspondente $\bar{p}$. Digamos que x denote o número de gerentes da amostra que concluíram o programa de treinamento gerencial. Os dados da Tabela 7.2 mostram que $x = 19$. Desse modo, com um tamanho de amostra $n = 30$, a proporção da amostra é

$$\bar{p} = \frac{x}{n} = \frac{19}{30} = 0{,}63$$

Ao fazer os cálculos anteriores, executamos o procedimento estatístico denominado *estimação por ponto*. Referimo-nos à média amostral $\bar{x}$ como o **estimador por ponto** da média da população m, ao desvio padrão da amostra s como o estimador por ponto do desvio padrão σ da população, e à proporção da

amostra $\bar{p}$ como o estimador por ponto da proporção p da população. O valor numérico de $\bar{x}$, s ou $\bar{p}$ chama-se **estimação por ponto**. Desse modo, em relação à amostra aleatória simples de 30 gerentes da EAI apresentada na Tabela 7.2, US$ 51.814 é a estimação por ponto de μ, US$ 3.348 é a estimação por ponto de σ e 0,63 é a estimação por ponto de p. A Tabela 7.3 resume os resultados amostrais e compara as estimações por ponto com os valores reais dos parâmetros populacionais.

Tabela 7.3 Resumo das estimações por ponto obtidas de uma amostra aleatória simples de 30 gerentes da EAI

Parâmetro Populacional	Valor do Parâmetro	Estimador por Ponto	Estimativa por Ponto
μ = Salário médio anual da população	US$ 51.800	$\bar{x}$ = Salário médio anual da amostra	US$ 51.814
σ = Desvio padrão do salário anual da população	US$ 4.000	s = Desvio padrão do salário anual da amostra	US$ 3.348
p = Proporção da população que concluiu o programa de treinamento gerencial	0,60	$\bar{p}$ = Proporção da população que concluiu o programa de treinamento gerencial	0,63

Como está claro na Tabela 7.3, a estimação por ponto difere bastante dos parâmetros populacionais correspondentes. Essa diferença deve ser esperada porque é usada uma amostra, não um censo da população inteira, para desenvolver as estimações por ponto. No próximo capítulo, mostraremos como construir uma estimação de intervalo para obtermos informações sobre quão próxima uma estimação por ponto está do parâmetro populacional.

Exercícios

Métodos

AUTOTESTE

11. Os dados a seguir são de uma amostra aleatória simples:

    5 8 10 7 10 14

    a. Qual é a estimação por ponto da média da população?
    b. Qual é a estimação por ponto do desvio padrão da população?

12. Uma pergunta de uma pesquisa realizada com uma amostra de 150 indivíduos produziu 75 respostas "Sim", 55 respostas "Não" e 20 "Sem Opinião".

    a. Qual é a estimação por ponto da proporção da população que respondeu Sim?
    b. Qual é a estimação por ponto da proporção da população que respondeu Não?

Aplicações

AUTOTESTE

13. Uma amostra aleatória simples dos dados de cinco meses de vendas forneceu a seguinte informação:

Mês	1	2	3	4	5
Unidades Vendidas	94	100	85	94	92

    a. Desenvolva uma estimação por ponto do número médio de unidades da população vendidas por mês.
    b. Desenvolva a estimação por ponto do desvio padrão da população.

ARQUIVO DA INTERNET
Mutual Fund

14. A *Business Week* publicou informações sobre 283 fundos mútuos de ações (*Business Week*, 26 de janeiro de 2004). Uma amostra de 40 desses fundos encontra-se no conjunto de dados (*data set*) MutualFund. Use o conjunto de dados para responder às seguintes questões:

    a. Desenvolva uma estimação por ponto da proporção dos fundos de ações da *Business Week* que são *load funds*.[4]

[4] NT: *Load funds* – Fundos mútuos com encargos (economia).

b. Desenvolva uma estimação por ponto da proporção de fundos que são classificados como investimentos de alto risco.
c. Desenvolva uma estimação por ponto da proporção de fundos que têm uma avaliação abaixo da média.

15. A *Appliance Magazine* publicou estimações da expectativa de durabilidade de aparelhos domésticos (*USA Today*, 5 de setembro de 2000). Uma amostra aleatória simples de dez aparelhos de videocassete (VCRs) apresenta os seguintes tempos de vida útil em termos de anos:

6,5 8,0 6,2 7,4 7,0 8,4 9,5 4,6 5,0 7,4

a. Desenvolva uma estimação por ponto da expectativa de durabilidade média da população de VCRs.
b. Desenvolva uma estimação por ponto do desvio padrão da expectativa de durabilidade média da população de VCRs.

16. Uma amostra de 50 empresas do grupo *Fortune* 500 (*Fortune*, 14 de abril de 2003) mostrou que cinco estavam sediadas em Nova York, seis na Califórnia, duas em Minnesota e uma em Wiconsin.

a. Desenvolva uma estimativa da proporção de empresas do grupo *Fortune* 500 sediadas em Nova York.
b. Desenvolva uma estimativa do número de empresas do grupo *Fortune* 500 sediadas em Minnesota.
c. Desenvolva uma estimativa da proporção de empresas do grupo *Fortune* 500 que não estão sediadas nesses quatro estados.

17. Uma pesquisa de opinião realizada pela Louis Harris ouviu 1.008 adultos para saber o que as pessoas pensavam sobre a economia (*Business Week*, 7 de agosto de 2000). As respostas foram as seguintes:

595 adultos	A economia está crescendo.
332 adultos	A economia permanece mais ou menos estagnada.
81 adultos	A economia está se retraindo.

Desenvolva uma estimação por ponto dos seguintes parâmetros populacionais.

a. A proporção de todos os adultos que acham que a economia está crescendo.
b. A proporção de todos os adultos que acham que a economia está mais ou menos estagnada.
c. A proporção de todos os adultos que acham que a economia está se retraindo.

7.4 INTRODUÇÃO ÀS DISTRIBUIÇÕES AMOSTRAIS

Na seção anterior, dissemos que a média da amostra $\bar{x}$ é o estimador por ponto da média populacional μ, e que a proporção da amostra $\bar{p}$ é o estimador por ponto da proporção da população p. Em relação à amostra aleatória simples de 30 gerentes da EAI, apresentada na Tabela 7.2, a estimação por ponto de μ é $\bar{x}$ = US$ 51.814,00 e a estimação por ponto de p é $\bar{p}$ 0,63. Suponha que selecionemos outra amostra aleatória simples de 30 gerentes da EAI e obtenhamos as seguintes estimações por ponto:

$$\text{Média da amostra } \bar{x} = \text{US\$ } 52.670$$
$$\text{Proporção da amostra } \bar{p} = 0{,}70$$

Observe que foram obtidos diferentes valores de $\bar{x}$ e de $\bar{p}$. De fato, não se pode esperar que uma segunda amostra aleatória simples de 30 gerentes da EAI produza as mesmas estimações por ponto que a primeira amostra.

Suponha agora que repetimos o processo de selecionar uma amostra aleatória simples de 30 gerentes da EAI diversas vezes, calculando a cada vez os valores de $\bar{x}$ e de $\bar{p}$. A Tabela 7.4 contém uma parte dos resultados obtidos para as 500 amostras aleatórias simples, e a Tabela 7.5 fornece as distribuições de freqüência e de freqüência relativa dos 500 valores de $\bar{x}$. A Figura 7.1 apresenta o histograma de freqüência relativa dos valores de $\bar{x}$.

No Capítulo 5, definimos uma variável aleatória como uma descrição numérica do resultado de um experimento. Se considerarmos que o processo de escolher uma amostra aleatória simples é um experimento, a média amostral $\bar{x}$ é uma descrição numérica do resultado do experimento. Desse modo, a média amostral $\bar{x}$ é uma variável aleatória. Conseqüentemente, à semelhança do que ocorre com qualquer variável aleatória, $\bar{x}$ tem um valor médio ou esperado, um desvio padrão e uma distribuição de probabilidade. Uma vez que os diversos valores possíveis de $\bar{x}$ resultam de diferentes amostras aleatórias simples, a distribuição da probabilidade de $\bar{x}$ é chamada **distribuição amostral** de $\bar{x}$. Conhecer essa distribuição amos-

tral e suas propriedades nos possibilitará fazer afirmações a respeito de quão próxima a média da amostra $\bar{x}$ está da média da população μ.

Tabela 7.4 Valores de $\bar{x}$ e $\bar{p}$ em 500 amostras aleatórias simples de 30 gerentes da EAI

Número da Amostra	Média da Amostra ($\bar{x}$)	Proporção da Amostra ($\bar{p}$)
1	51.814	0,63
2	52.670	0,70
3	51.780	0,67
4	51.588	0,53
.	.	.
.	.	.
.	.	.
500	51.752	0,50

Retornemos à Figura 7.1. Precisaríamos enumerar cada amostra possível de 30 gerentes e calcular cada média amostral para determinar de maneira completa a distribuição amostral de $\bar{x}$. Entretanto, o histograma de 500 valores de $\bar{x}$ fornece uma aproximação dessa distribuição amostral. Pela aproximação, observamos que a distribuição tem a forma de sino. Notamos que a maior concentração dos valores de $\bar{x}$ e a média dos 500 valores de $\bar{x}$ estão próximas da média populacional μ = US\$ 51.800. Descreveremos as propriedades das distribuições amostrais de $\bar{x}$ mais detalhadamente na próxima seção.

Os 500 valores da proporção da amostra $\bar{p}$ são sintetizados pelo histograma de freqüência relativa da Figura 7.2. Assim como ocorre com $\bar{x}$, $\bar{p}$ é uma variável aleatória. Se toda amostra possível de tamanho 30 fosse selecionada da população, e se um valor de $\bar{p}$ fosse calculado para cada amostra, a distribuição de probabilidade resultante seria a distribuição amostral de $\bar{p}$. O histograma de freqüência relativa dos 500 valores da amostra apresentado na Figura 7.2 nos dá uma idéia geral da aparência da distribuição amostral de $\bar{p}$.

Na prática, selecionamos somente uma amostra aleatória simples da população. Repetimos o processo de amostragem 500 vezes nesta seção simplesmente para ilustrar que muitas amostras diferentes são possíveis e que as diferentes amostras geram uma grande variedade de valores para as estatísticas da amostra $\bar{x}$ e $\bar{p}$. A distribuição de probabilidade de qualquer estatística amostral em particular é denominada distribuição amostral. Na Seção 7.5, apontaremos as características da distribuição amostral de $\bar{x}$. Na Seção 7.6, mostraremos as características da distribuição amostral de $\bar{p}$.

Tabela 7.5 Distribuição da freqüência de $\bar{x}$ em 500 amostras aleatórias simples de 30 gerentes da EAI

Salário Anual Médio ($)	Freqüência	Freqüência Relativa
49.500,00–49.999,99	2	0,004
50.000,00–50.499,99	16	0,032
50.500,00–50.999,99	52	0,104
51.000,00–51.499,99	101	0,202
51.500,00–51.999,99	133	0,266
52.000,00–52.499,99	110	0,220
52.500,00–52.999,99	54	0,108
53.000,00–53.499,99	26	0,052
53.500,00–53.999,99	6	0,012
Totais	500	1.000

Figura 7.1 Histograma da freqüência relativa dos valores de $\bar{x}$ em 500 amostras aleatórias simples com tamanho 30 cada uma

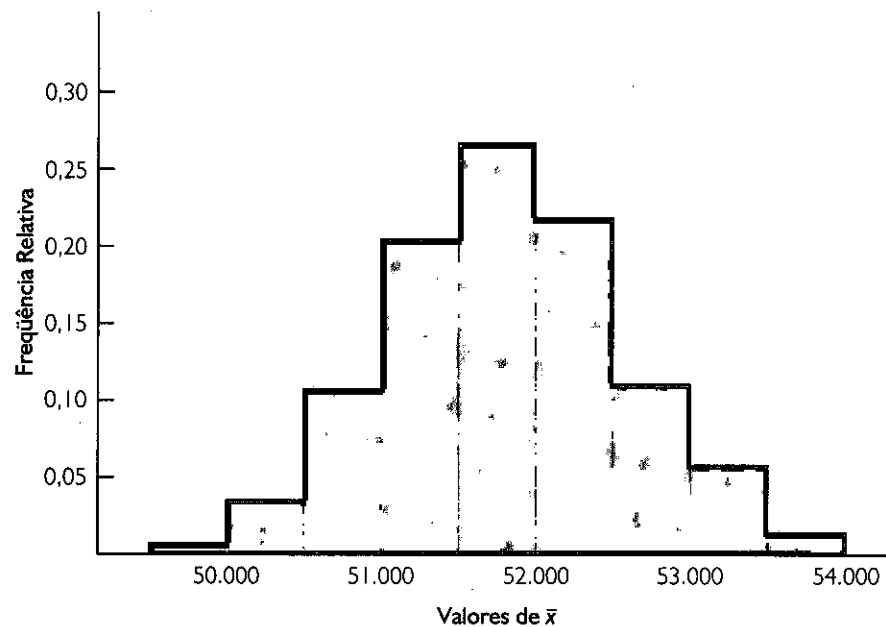
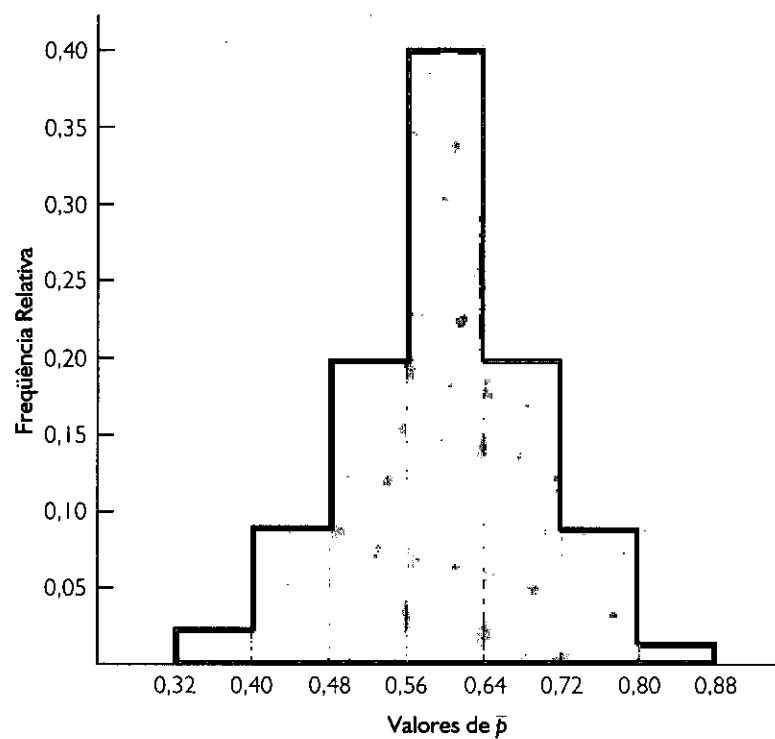



Figura 7.2 Histograma da freqüência relativa dos valores de $\bar{p}$ em 500 amostras aleatórias simples com tamanho 30 cada uma

7.5 DISTRIBUIÇÃO AMOSTRAL DE $\bar{x}$

Na seção anterior, dissemos que a média da amostra $\bar{x}$ é uma variável aleatória e que sua distribuição de probabilidade é chamada distribuição amostral de $\bar{x}$.

DISTRIBUIÇÃO AMOSTRAL DE $\bar{x}$

A distribuição amostral de $\bar{x}$ é a distribuição de probabilidade de todos os valores possíveis da média amostral $\bar{x}$.

Esta seção descreve as propriedades da distribuição amostral de $\bar{x}$. Exatamente como ocorre com outras distribuições de probabilidade que estudamos, a distribuição amostral de $\bar{x}$ tem um valor esperado (ou média), um desvio padrão e um formato, ou forma, característico. Vamos iniciar considerando a média de todos os valores possíveis de $\bar{x}$, à qual nos referimos como valor esperado de $\bar{x}$.

Valor Esperado de $\bar{x}$

No problema de amostragem da EAI, vimos que diferentes amostras aleatórias simples resultam em uma série de valores correspondentes à média amostral $\bar{x}$. Como são possíveis muitos valores diferentes da variável aleatória $\bar{x}$, freqüentemente o que nos interessa é a média de todos os possíveis valores de $\bar{x}$ que podem ser gerados pelas várias amostras aleatórias simples. A média da variável aleatória $\bar{x}$ é o valor esperado de $\bar{x}$. Admitamos que $E(\bar{x})$ representa o valor esperado de $\bar{x}$ e μ representa a média da população da qual estamos selecionando uma amostra aleatória simples. Podemos demonstrar que, quando se trata de uma amostragem aleatória simples, $E(\bar{x})$ e μ são iguais.

VALOR ESPERADO DE $\bar{x}$

$$E(\bar{x}) = \mu \tag{7.1}$$

em que

$E(\bar{x})$ = o valor esperado de $\bar{x}$

μ = a média da população

O valor esperado de $\bar{x}$ é igual à média da população da qual a amostra é selecionada.

Esse resultado mostra que, quando se trata de uma amostragem aleatória simples, o valor esperado (ou média) da distribuição amostral de $\bar{x}$ é igual à média da população. Na Seção 7.1, vimos que o salário anual médio da população de gerentes da EAI é μ = US\$ 51.800. Desse modo, de acordo com a Equação 7.1, a média de todas as médias amostrais possíveis no estudo da EAI é também US\$ 51.800.

Quando o valor esperado de um estimador por ponto for igual ao parâmetro populacional, dizemos que o estimador por ponto é **sem viés**. Assim, a Equação 7.1 mostra que $\bar{x}$ é um estimador sem viés da média populacional μ.

Desvio Padrão de $\bar{x}$

Vamos definir o desvio padrão da distribuição amostral de $\bar{x}$. Usaremos a seguinte notação:

$\sigma_{\bar{x}}$ = o desvio padrão de $\bar{x}$
σ = o desvio padrão da população
n = o tamanho da amostra
N = o tamanho da população

Pode-se demonstrar que, quando se trata de amostragem aleatória simples, o desvio padrão de $\bar{x}$ depende de a população ser finita ou infinita. As duas expressões para o desvio padrão de $\bar{x}$ são as seguintes:

DESVIO PADRÃO DE $\bar{x}$

População Finita	*População Infinita*	
$\sigma_{\bar{x}} = \sqrt{\frac{N-n}{N-1}}\left(\frac{\sigma}{\sqrt{n}}\right)$	$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$	(7.2)

Ao comparar as duas fórmulas apresentadas na Equação 7.2, vemos que o fator $\sqrt{(N-n)/(N-1)}$ é necessário para o caso de a população ser finita, mas não para o caso da população infinita. Esse fator comumente é chamado **fator de correção para populações finitas**. Em muitas situações práticas de amostragem, descobrimos que a população, não obstante ser finita, é "grande", ao passo que o tamanho da amostra é relativamente "pequeno". Nesses casos, o fator de correção para populações finitas $\sqrt{(N-n)/(N-1)}$ está próximo de 1. Conseqüentemente, a diferença entre os valores do desvio padrão de $\bar{x}$ para os casos de populações finitas e infinitas torna-se desprezível. Então, $\sigma_{\bar{x}} = \sigma/\sqrt{n}$ torna-se uma boa aproximação ao desvio padrão de $\bar{x}$, embora a população seja finita. Essa observação leva à seguinte diretriz geral, ou método prático, de se calcular o desvio padrão de $\bar{x}$.

USE A SEGUINTE EXPRESSÃO PARA CALCULAR O DESVIO PADRÃO DE $\bar{x}$

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \tag{7.3}$$

sempre que

1. A população for infinita; ou
2. A população for finita *e* o tamanho da amostra for menor ou igual a 5% do tamanho da população; ou seja, $n/N \leq 0{,}05$.

O Problema 21 demonstra o seguinte: quando $n/N \leq 0{,}05$, o fator de correção para populações finitas tem pouco efeito sobre o valor de $\sigma_{\bar{x}}$.

O termo erro padrão é usado quando queremos nos referir ao desvio padrão de um estimador por ponto.

Nos casos em que $n/N > 0{,}05$, a versão para populações finitas da Equação 7.2 deve ser usada no cálculo de $\sigma_{\bar{x}}$. A menos que seja indicado o contrário, ao longo de todo o livro presumiremos que o tamanho da população seja "grande", $n/N \leq 0{,}05$, e a Equação 7.3 pode ser usada para calcular $\sigma_{\bar{x}}$.

Para calcular $\sigma_{\bar{x}}$, precisamos conhecer σ, que é o desvio padrão da população. Para enfatizarmos ainda mais a diferença entre $\sigma_{\bar{x}}$ e σ, referimo-nos ao desvio padrão de $\bar{x}$, $\sigma_{\bar{x}}$, como o **erro padrão** da média. Em geral, o termo *erro padrão* refere-se ao desvio padrão de um estimador por ponto. Posteriormente, veremos que o valor do erro padrão da média é útil para determinarmos quão distante a média amostral pode estar da média da população. Retornemos agora ao exemplo da EAI e calculemos o erro padrão da média associada às amostras aleatórias simples de 30 gerentes da EAI.

Na Seção 7.1, vimos que o desvio padrão dos salários anuais da população de 2.500 gerentes da EAI é $\sigma = 4.000$. Nesse caso, a população é finita, com $N = 2.500$. Entretanto, com um tamanho de amostra igual a 30, temos $n/N = 30/2.500 = 0{,}012$. Uma vez que o tamanho da amostra é menor que 5% do tamanho da população, podemos ignorar o fator de correção para populações finitas e usar a Equação (7.3) para calcular o erro padrão.

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{4.000}{\sqrt{30}} = 730{,}3$$

Forma da Distribuição Amostral de $\bar{x}$

Os resultados anteriores referentes ao valor esperado e ao desvio padrão da distribuição amostral de $\bar{x}$ são aplicáveis a qualquer população. A etapa final do processo de identificação das características da distribuição amostral de $\bar{x}$ é determinar o formato, ou forma, da distribuição amostral. Consideraremos dois casos: (1) a população tem uma distribuição normal; e (2) a população não tem uma distribuição normal.

A população tem uma distribuição normal Em muitas situações, é razoável supormos que a população da qual selecionamos uma amostra aleatória simples em uma distribuição normal, ou aproximadamente normal. Quando a população tem uma distribuição normal, a distribuição amostral de $\bar{x}$ está normalmente distribuída para qualquer tamanho de amostra.

A população não tem uma distribuição normal Quando a população da qual selecionamos uma amostra aleatória simples não tem uma distribuição normal, o **teorema do limite central** é útil para identificarmos a forma da distribuição amostral de $\bar{x}$. Uma definição do teorema do limite central, quando ele se aplica à distribuição amostral de $\bar{x}$, é a seguinte:

TEOREMA DO LIMITE CENTRAL
Ao selecionar amostras aleatórias simples de tamanho n de uma população, podemos aproximar a distribuição amostral da média da amostra $\bar{x}$ por meio de uma *distribuição normal* à medida que o tamanho da amostra se torna maior.

A Figura 7.3 mostra como o teorema do limite central funciona em relação a três populações diferentes: cada coluna refere-se a uma das populações. O painel superior da figura mostra que nenhuma das populações está normalmente distribuída. A população I segue uma distribuição uniforme. A população II, muitas vezes, é chamada distribuição "orelha-de-coelho". Ela é simétrica, mas os valores mais prováveis situam-se nas extremidades (caudas). A população III tem uma forma similar à da distribuição exponencial; ela tem uma inflexão à direita.

Os três painéis da parte inferior da Figura 7.3 mostram a forma da distribuição amostral correspondente a amostras de tamanho $n = 2$, $n = 5$ e $n = 30$. Quando a amostra tem tamanho 2, notamos que a forma de cada distribuição amostral é diferente da forma da distribuição populacional correspondente. Em relação a amostras de tamanho 5, notamos que a forma da distribuição amostral referente às populações I e II começa a demonstrar certa similaridade com a forma da distribuição normal. Não obstante a forma da distribuição amostral da população III começar a demonstrar similaridade com a forma de uma distribuição normal, ainda há certa inflexão à direita. Finalmente, para amostras de tamanho 30, as formas de cada uma das três distribuições amostrais são aproximadamente normais.

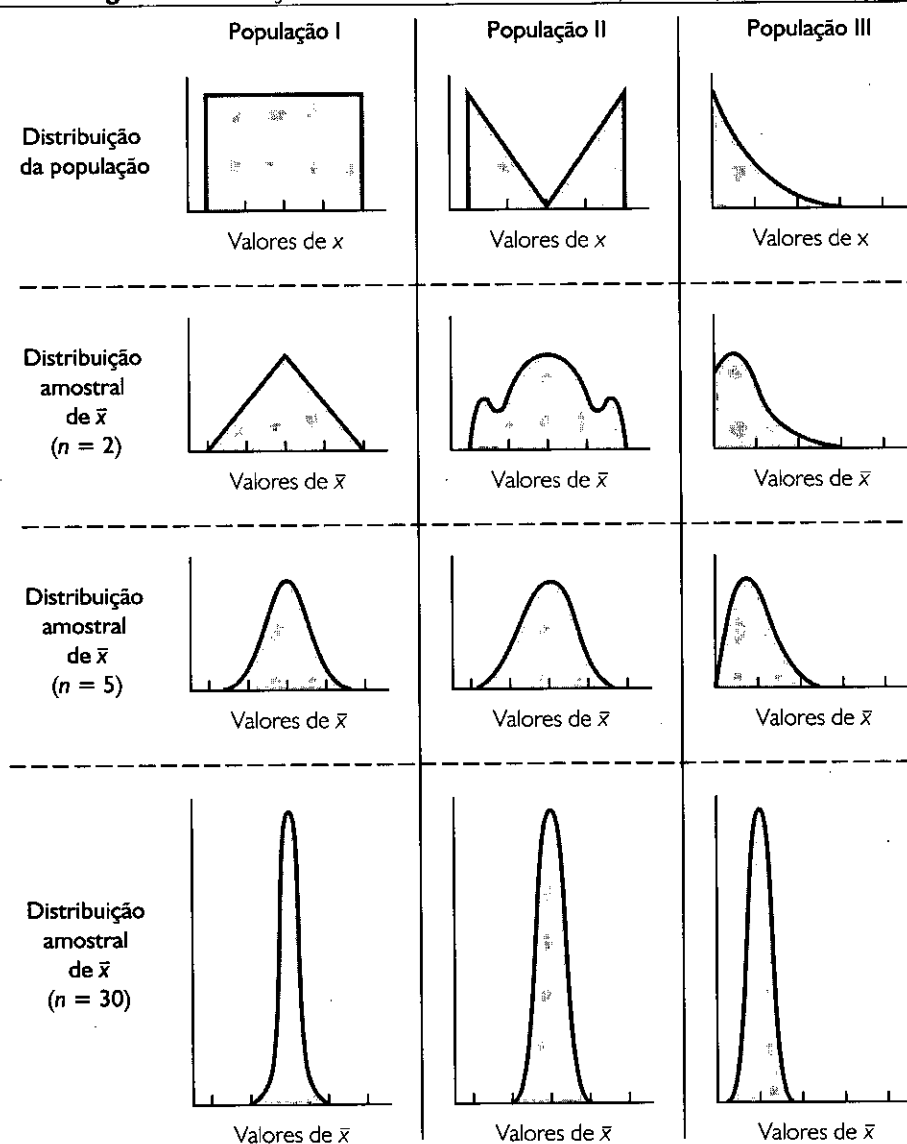
Figura 7.3 Ilustração do teorema do limite central para três populações

População I
População II
População III

Distribuição da população
Valores de x
Valores de x
Valores de x

Distribuição amostral de $\bar{x}$ ($n = 2$)
Valores de $\bar{x}$
Valores de $\bar{x}$
Valores de $\bar{x}$

Distribuição amostral de $\bar{x}$ ($n = 5$)
Valores de $\bar{x}$
Valores de $\bar{x}$
Valores de $\bar{x}$

Distribuição amostral de $\bar{x}$ ($n = 30$)
Valores de $\bar{x}$
Valores de $\bar{x}$
Valores de $\bar{x}$

Do ponto de vista profissional, freqüentemente queremos saber quão grande o tamanho da amostra precisa ser antes de o teorema do limite central aplicar-se e podermos presumir que a forma da distribuição amostral seja aproximadamente normal. Pesquisadores estatísticos investigaram essa questão estudando a distribuição amostral de $\bar{x}$ para uma grande variedade de populações e de tamanhos de amostra. A prática geral da estatística é supor que, para a maioria das aplicações, a distribuição amostral de $\bar{x}$ pode ser aproximada por meio de uma distribuição normal sempre que a amostra tiver tamanho 30 ou mais. Nos casos em que a população tem uma inflexão elevada ou existam pontos fora da curva, podem ser necessárias amostras de tamanho 50. Finalmente, se a população for discreta, o tamanho de amostra necessário a uma distribuição normal dependerá muitas vezes da proporção da população. Falaremos mais sobre esse assunto quando discutirmos a distribuição amostral de $\bar{p}$ na Seção 7.6.

Distribuição Amostral de $\bar{x}$ para o Problema da EAI

Retornemos ao problema da EAI, na parte em que mostramos anteriormente que $E(\bar{x})$ = US\$ 51.800 e $\sigma_{\bar{x}} = 730{,}3$. Neste ponto, não temos nenhuma informação sobre a distribuição da população; ela pode estar distribuída normalmente ou não. Se a população tem uma distribuição normal, a distribuição amostral de $\bar{x}$ está normalmente distribuída. Se a população não tem uma distribuição normal, a amostra aleatória simples de 30 gerentes e o teorema do limite central nos possibilitam concluir que a distribuição amostral de $\bar{x}$ pode ser aproximada por meio de uma distribuição normal. Em qualquer dos casos, sentimo-nos à vontade em prosseguir com a conclusão de que a distribuição amostral de $\bar{x}$ pode ser descrita pela distribuição normal mostrada na Figura 7.4.

Valor Prático da Distribuição Amostral de $\bar{x}$

Sempre que uma amostra aleatória simples é selecionada e o valor da média da amostra $\bar{x}$ é usado para estimar o valor da média da população μ, não podemos esperar que a média da amostra seja exatamente igual à média da população. A razão prática pela qual estamos interessados na distribuição amostral de $\bar{x}$ é que ela pode ser usada para fornecer informações probabilísticas a respeito da diferença entre a média da amostra e a média da população. Para demonstrar esse uso, retornemos ao problema da EAI.

Suponha que o diretor de pessoal acredite que a média da amostra venha a ser uma estimativa aceitável da média da população se essa média da amostra estiver dentro de US$ 500 da média da população. Entretanto, não é possível garantir que a média da amostra estará dentro de US$ 500 da média da população. De fato, a Tabela 7.5 e a Figura 7.1 mostram que algumas das 500 médias da amostra diferiam em mais de US$ 2 mil da média da população. Então, precisamos pensar no pedido do diretor de pessoal em termos probabilísticos. Ou seja, o diretor de pessoal preocupa-se com a seguinte questão: qual é a probabilidade de a média da amostra, calculada usando-se uma amostra aleatória simples de 30 gerentes da EAI, estar dentro de US$ 500 da média da população?

Já que identificamos as propriedades da distribuição amostral de $\bar{x}$ (veja a Figura 7.4), usaremos essa distribuição para responder à questão de probabilidade. Consulte a distribuição amostral de $\bar{x}$ apresentada novamente na Figura 7.5. Com uma média populacional de US\$ 51.800, o diretor de pessoal quer saber qual é a probabilidade de $\bar{x}$ estar entre US\$ 51.300 e US\$ 52.300. Essa probabilidade é dada pela área com sombreamento mais escuro da distribuição amostral apresentada na Figura 7.5. Uma vez que a distribuição amostral está normalmente distribuída, com a média de 51.800 e erro padrão da média igual a 730,3, podemos usar a tabela de áreas da distribuição normal padrão. Para $\bar{x} = 51.300$, temos:

$$z = \frac{51.300 - 51.800}{730{,}3} = -0{,}68$$

Figura 7.4 Distribuição amostral de $\bar{x}$ do salário médio anual de uma amostra aleatória simples de 30 gerentes da EAI

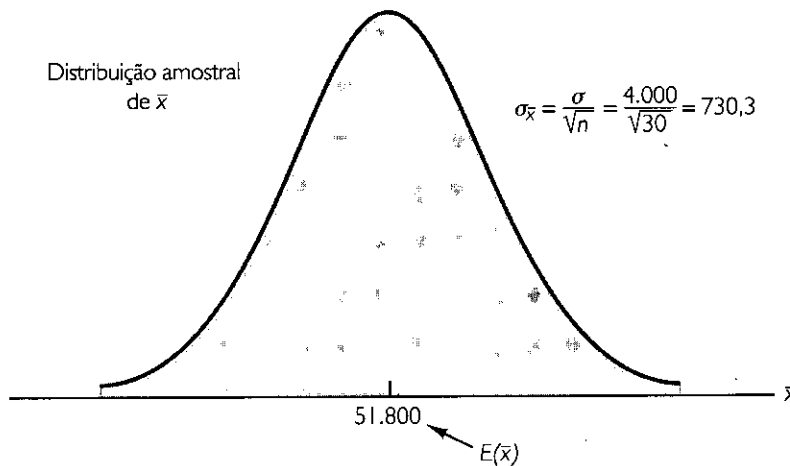
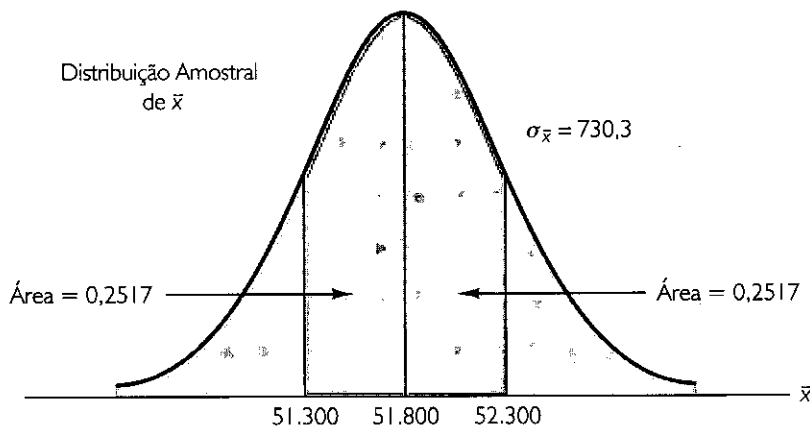



Figura 7.5 A probabilidade de uma média amostral estar dentro de US$ 500 da média da população

Consultando a tabela de distribuição normal padrão, verificamos que a área entre $z = 0$ e $z = -0,68$ é 0,2517. Cálculos similares para $\bar{x} = 52.300$ mostram que a área entre $z = 0$ e $z = +0,68$ corresponde a 0,2517. Desse modo, a probabilidade de o valor da média da amostra estar entre 51.300 e 52.300 é 0,2517 + 0,2517 = 0,5034.

A distribuição amostral de $\bar{x}$ pode ser usada para produzir informações probabilísticas a respeito de quão próxima a média amostral $\bar{x}$ está da média populacional μ.

Os cálculos anteriores revelam que uma amostra aleatória simples de 30 gerentes da EAI tem uma probabilidade de 0,5034 de produzir uma média amostral $\bar{x}$ que esteja dentro de US$ 500 da média da população. Assim, há a probabilidade de $1 - 0,5034 = 0,4966$ de a diferença entre $\bar{x}$ e μ = US$ 51.800 ser maior que US$ 500. Em outras palavras, uma amostra aleatória simples de 30 gerentes da EAI tem aproximadamente 50–50 de chances de produzir uma média amostral dentro dos US$ 500 admissíveis. Talvez um tamanho de amostra maior deva ser considerado. Vamos explorar essa possibilidade considerando a relação entre o tamanho da amostra e a distribuição amostral de $\bar{x}$.

Relação entre o Tamanho da Amostra e a Distribuição Amostral de $\bar{x}$

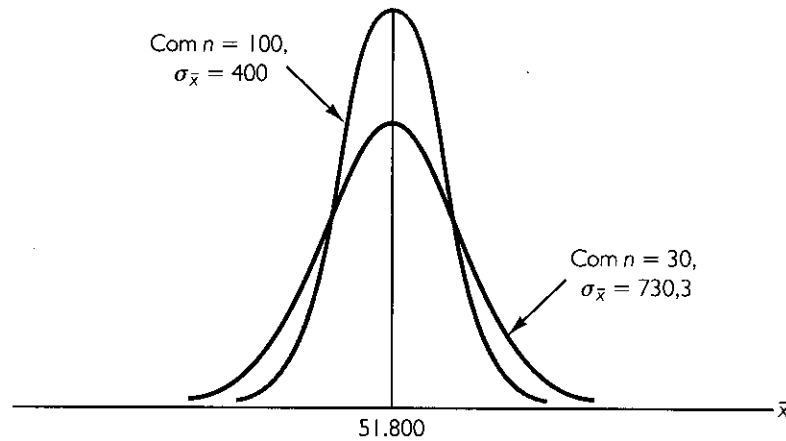
Suponha que, no problema de amostragem da EAI, selecionemos uma amostra aleatória simples de cem gerentes em vez dos 30 considerados a princípio. Intuitivamente, poderia parecer que, em decorrência da maior quantidade de dados oferecidos pelo maior tamanho de amostra, a média amostral baseada em

$n = 100$ nos daria uma estimativa melhor da média populacional que a média amostral baseada em $n = 30$. Para ver melhor, consideremos a relação entre o tamanho da amostra e a distribuição amostral de $\bar{x}$.

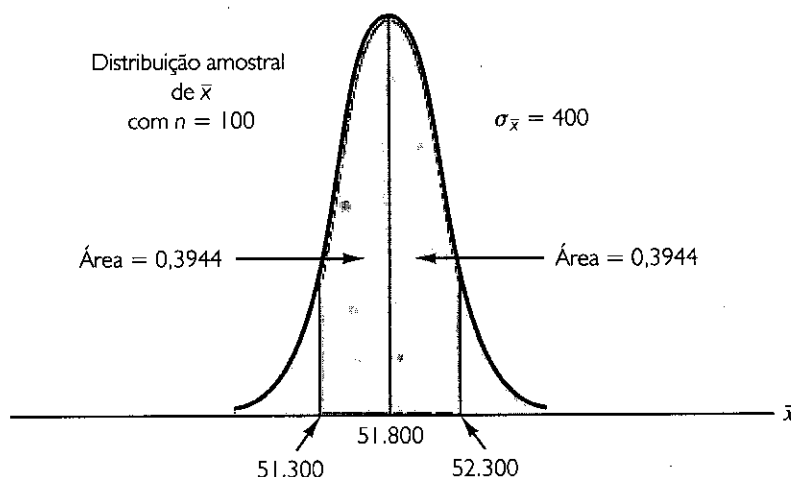
Primeiramente, observe que $E(\bar{x}) = \mu$, independentemente do tamanho da amostra. Assim, a média de todos os valores possíveis de $\bar{x}$ é igual à média da população μ, independentemente do tamanho da amostra n. Note, entretanto, que o erro padrão da média, $\sigma_{\bar{x}} = \sigma/\sqrt{n}$, está relacionado com a raiz quadrada do tamanho da amostra. Sempre que o tamanho da amostra for aumentado, o erro padrão da média $\sigma_{\bar{x}}$ decresce. Com $n = 30$, o erro padrão da média relativo ao problema da EAI é 730,3. Porém, com o aumento do tamanho da amostra para $n = 100$, o erro padrão da média decresce para

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{4.000}{\sqrt{100}} = 400$$

As distribuições amostrais de $\bar{x}$, com $n = 30$ e $n = 100$, são mostradas na Figura 7.6. Desde que a distribuição amostral com $n = 100$ tenha um erro padrão menor, os valores de $\bar{x}$ têm menos variação e tendem a aproximar-se mais da média da população que os valores de $\bar{x}$ com $n = 30$.

Figura 7.6 Comparação das distribuições amostrais de $\bar{x}$ para amostras aleatórias simples de $n = 30$ e $n = 100$ gerentes da EAI

Podemos usar a distribuição amostral $\bar{x}$ para o caso em que $n = 100$ para comparar a probabilidade de uma amostra aleatória simples de 100 gerentes da EAI produzir uma média amostral que esteja dentro dos US$ 500 da média da população. Uma vez que a distribuição amostral é normal, com uma média igual a 51.800 e erro padrão igual a 400, podemos usar a tabela de distribuição normal padrão para encontrar a área ou a probabilidade. Para $\bar{x} = 51.300$ (Figura 7.7), temos

$$z = \frac{51.300 - 51.800}{400} = -1,25$$

Figura 7.7 A probabilidade de uma média amostral estar dentro dos US$ 500 da média da população quando se usa uma amostra aleatória simples de 100 gerentes da EAI

Consultando a tabela de distribuição normal padrão de probabilidade, verificamos que a área entre $z = 0$ e $z = -1{,}25$ é 0,3944. Com um cálculo similar para $\bar{x} = 52.300$, vemos que a probabilidade de o valor da média da amostra estar entre 51.300 e 52.300 é 0,3944 + 0,3944 = 0,7888. Desse modo, ao aumentarmos o tamanho da amostra de 30 para 100 gerentes da EAI, elevamos também a probabilidade de obter uma média amostral dentro dos US$ 500 da média da população, de 0,5034 para 0,7888.

O ponto importante nesta discussão é que, à medida que se aumenta o tamanho da amostra, o erro padrão da média diminui. Conseqüentemente, quanto maior o tamanho da amostra, maior a probabilidade de a média da amostra estar dentro de uma distância específica da média da população.

NOTAS E COMENTÁRIOS

1. Ao apresentar a distribuição amostral de $\bar{x}$ relativa ao problema da EAI, recorremos ao fato de que a média populacional $\mu = 51.800$ e o desvio padrão da população $\sigma = 4.000$ eram conhecidos. Entretanto, geralmente os valores da média populacional μ e o desvio padrão da população σ, que são necessários para determinar a distribuição amostral de $\bar{x}$, são desconhecidos. No Capítulo 8, mostraremos como a média da amostra $\bar{x}$ e o desvio padrão da amostra s são usados quando μ e σ são desconhecidos.
2. A demonstração teórica do teorema do limite central requer observações independentes na amostra. Essa condição é satisfeita para populações infinitas e finitas em que a amostragem é feita com substituição. Embora o teorema do limite central não lide diretamente com amostragens sem substituição de populações finitas, a prática geral da estatística aplica as conclusões do teorema do limite central a essa situação quando o tamanho da população é grande.

Exercícios

Métodos

18. A média de uma população é 200 e seu desvio padrão é 50. Uma amostra aleatória simples de tamanho 100 será tomada e a média amostral $\bar{x}$ será usada para estimar a média da população.

    a. Qual é o valor esperado de $\bar{x}$?
    b. Qual é o desvio padrão de $\bar{x}$?
    c. Apresente a distribuição amostral de $\bar{x}$.
    d. O que a distribuição amostral de $\bar{x}$ indica?

AUTOTESTE

19. A média de uma população é 200 e seu desvio padrão é 50. Suponha que uma amostra aleatória simples de tamanho 100 seja selecionada e que $\bar{x}$ seja usado para estimar μ.
   a. Qual é a probabilidade de a média da amostra estar dentro de ± 5 da média da população?
   b. Qual é a probabilidade de a média da amostra estar dentro de ± 10 da média da população?
20. Suponha que o desvio padrão da população seja $\sigma = 25$. Calcule o erro padrão da média $\sigma_{\bar{x}}$, para tamanhos de amostra iguais a 50, 100, 150 e 200. O que se pode afirmar sobre o tamanho do erro padrão da média quando o tamanho da amostra for aumentado?
21. Suponha que uma amostra aleatória simples de tamanho 50 seja selecionada de uma população com $\sigma = 10$. Encontre o valor do erro padrão da média em cada um dos seguintes casos (use o fator de correção para populações finitas, se for o caso).
   a. O tamanho da população é infinito.
   b. O tamanho da população é $N = 50.000$.
   c. O tamanho da população é $N = 5.000$.
   d. O tamanho da população é $N = 500$.

Aplicações

22. Consulte o problema de amostragem da EAI. Suponha que seja usada uma amostra aleatória simples de 60 gerentes.
   a. Trace um esboço da distribuição amostral de $\bar{x}$ quando são usadas amostras aleatórias simples de tamanhos 60.
   b. O que acontece com a distribuição amostral de $\bar{x}$ se forem usadas amostras aleatórias simples de tamanho 120?
   c. Qual afirmação genérica se pode fazer a respeito daquilo que acontece à distribuição amostral de $\bar{x}$ quando o tamanho da amostra for aumentado? Essa generalização parece lógica? Explique.

AUTOTESTE

23. No problema de amostragem da EAI (veja a Figura 7.5), mostramos que para $n = 30$ havia a probabilidade de 0,5034 de obtermos uma média amostral dentro de ± US$ 500 da média da população.
   a. Qual é a probabilidade de $\bar{x}$ estar dentro de US$ 500 da média da população, se for usado um tamanho de amostra igual a 60?
   b. Responda ao item (a) considerando uma amostra com tamanho 120.
24. O custo médio do ensino nas universidades públicas norte-americanas é US$ 4.260 por ano (*St. Petersburg Times*, 11 de dezembro de 2002). Use esse valor como média populacional e considere que o desvio padrão da população é σ = US$ 900. Suponha que uma amostra aleatória de 50 universidades públicas seja selecionada.
   a. Apresente a distribuição amostral de $\bar{x}$ em que $\bar{x}$ é a média amostral do custo de ensino nas 50 universidades.
   b. Qual é a probabilidade de a amostra aleatória simples produzir uma média amostral que se situe dentro dos US$ 250 da média populacional?
   c. Qual é a probabilidade de a amostra aleatória simples produzir uma média amostral que se situe dentro dos US$ 100 da média populacional?
25. O *College Board American College Testing Program* divulgou que a média populacional das pontuações nos exames SAT é $\mu = 1.020$ (*The World Almanac 2003*). Considere que o desvio padrão da população seja $\sigma = 100$.
   a. Qual é a probabilidade de uma amostra aleatória de 75 estudantes produzir uma média amostral de pontuações SAT dentro de 10 da média populacional?
   b. Qual é a probabilidade de uma amostra aleatória de 75 estudantes produzir uma média amostral de pontuação SAT dentro de 20 da média populacional?
26. O salário anual inicial médio de graduados com *major*[5] em marketing é US$ 34 mil (*Time*, 8 de maio de 2000). Suponha que o salário anual inicial médio da população de graduados com *major* em marketing seja $\mu = 34.000$ e o desvio padrão seja $\sigma = 2.000$.

[5] NT: *Major*: *Educ*. – Designa a área de estudo universitário na qual o estudante se especializa (Estados Unidos).

a. Qual é a probabilidade de uma amostra aleatória simples de graduados com *major* em marketing ter uma média amostral dentro de ± US$ 250 da média populacional correspondente a cada um dos seguintes tamanhos de amostra: 30, 50, 100, 200 e 400?

b. Qual é a vantagem de um tamanho maior de amostra quando se tenta estimar a média da população?

27. A *Business Week* realizou uma pesquisa de opinião de graduados dos 30 melhores programas de MBA (*Business Week*, 22 de setembro de 2003). A pesquisa revelou que o salário anual médio de homens e mulheres, dez anos após a graduação, eram US$ 168 mil e US$ 117 mil, respectivamente. Suponha que o desvio padrão para os graduados seja US$ 40 mil, e para as graduadas seja US$ 25 mil.

a. Qual é a probabilidade de uma amostra aleatória simples de 40 homens, graduados produzir uma média amostral dentro dos US$ 10 mil da média populacional, US$ 168 mil?

b. Qual é a probabilidade de uma amostra aleatória simples de 40 mulheres, graduadas produzir uma média amostral dentro dos US$ 10 mil da média populacional, US$ 117 mil?

c. Em qual dos dois casos anteriores, item (a) ou item (b), temos maior probabilidade de obter uma estimativa amostral dentro dos US$ 10 mil da média populacional? Por quê?

d. Qual é a probabilidade de uma amostra aleatória simples de cem graduados, homens, produzir uma média amostral maior que US$ 4 mil abaixo da média populacional?

28. O custo médio anual dos seguros de automóvel é US$ 687 (*National Association of Insurance Commissioners*, janeiro de 2003). Use esse valor como média populacional e suponha que o desvio padrão da população seja σ = US$ 230. Considere uma amostra de 45 apólices de seguro de automóveis.

a. Apresente a distribuição amostral de $\bar{x}$, em que $\bar{x}$ é a média amostral do custo anual dos seguros de automóvel.

b. Qual é a probabilidade de a média amostral estar dentro dos US$ 100 da média populacional?

c. Qual é a probabilidade de a média amostral estar dentro dos US$ 25 da média populacional?

d. O que você recomendaria se uma seguradora quisesse a média amostral para estimar a média populacional dentro de ±US$ 25?

29. A revista *Money* divulgou que o preço médio de um galão de gasolina nos Estados Unidos durante o primeiro trimestre de 2001 era US$ 1,46 (*Money*, agosto de 2001). Suponha que o preço divulgado pela *Money* seja a média populacional e que o desvio padrão populacional seja σ = US$ 0,15.

a. Qual é a probabilidade de o preço médio de uma amostra de 30 postos de gasolina estar dentro dos US$ 0,03 da média populacional?

b. Qual é a probabilidade de o preço médio de uma amostra de 50 postos de gasolina estar dentro dos US$ 0,03 da média populacional?

c. Qual é a probabilidade de o preço médio de uma amostra de 100 postos de gasolina estar dentro dos US$ 0,03 da média populacional?

d. Você recomendaria um tamanho de amostra de 30, 50 ou 100 para obter, no mínimo, uma probabilidade de 0,95 de que a média amostral se situe dentro dos US$ 0,03 da média populacional?

30. Para estimar a idade média de uma população de 4 mil empregados, foi selecionada uma amostra aleatória simples de 40 empregados.

a. Você usaria o fator de correção para populações finitas ao calcular o erro padrão da média? Explique.

b. Se o desvio padrão da população é $\sigma = 8{,}2$ anos, calcule o erro padrão utilizando o fator de correção para populações finitas e sem utilizá-lo. Qual é o fundamento lógico para se ignorar o fator de correção para populações finitas sempre que $n/N \leq 0{,}05$?

c. Qual é a probabilidade de a média amostral de idade dos empregados estar dentro de ±2 anos da idade média da população?

7.6 DISTRIBUIÇÃO AMOSTRAL DE $\bar{p}$

A proporção amostral $\bar{p}$ é o estimador por ponto da proporção p da população. A fórmula para calcular a proporção amostral é:

$$\bar{p} = \frac{x}{n}$$

em que

x = o número de elementos contidos na amostra que possuem a característica de interesse.
n = o tamanho da amostra.

Conforme observamos na Seção 7.4, a proporção amostral $\bar{p}$ é uma variável aleatória e sua distribuição de probabilidade denomina-se distribuição amostral de $\bar{p}$.

DISTRIBUIÇÃO AMOSTRAL DE $\bar{p}$

A distribuição amostral de $\bar{p}$ é a distribuição de probabilidade de todos os valores possíveis da proporção amostral $\bar{p}$.

Para determinar quão próxima a proporção amostral $\bar{p}$ está da proporção populacional p, precisamos entender as propriedades da distribuição amostral de $\bar{p}$: o valor esperado de $\bar{p}$, o desvio padrão de $\bar{p}$ e a forma, ou formato, da distribuição amostral de $\bar{p}$.

Valor Esperado de $\bar{p}$

O valor esperado de $\bar{p}$, que é a média de todos os valores possíveis de $\bar{p}$, é igual à proporção populacional de p.

VALOR ESPERADO DE $\bar{p}$

$$E(\bar{p}) = p \qquad (7.4)$$

em que

$E(\bar{p})$ = o valor esperado de $\bar{p}$
p = a proporção populacional

Uma vez que $E(\bar{p}) = p$, $\bar{p}$ é um estimador sem viés de p. Lembre-se de que observamos na Seção 7.1 que $p = 0{,}60$ para a população da EAI, em que p é a proporção da população de gerentes que participaram do programa de treinamento gerencial da empresa. Desse modo, o valor esperado de $\bar{p}$ para o problema de amostragem da EAI é 0,60.

Desvio Padrão de $\bar{p}$

Exatamente como concluímos em relação ao desvio padrão de $\bar{p}$, o desvio padrão depende de a população ser finita ou infinita. As duas fórmulas para calcular o desvio padrão de $\bar{p}$ são as seguintes:

DESVIO PADRÃO DE $\bar{p}$

População Finita

$$\sigma_{\bar{p}} = \sqrt{\frac{N-n}{N-1}}\sqrt{\frac{p(1-p)}{n}}$$

População Infinita

$$\sigma_{\bar{p}} = \sqrt{\frac{p(1-p)}{n}} \qquad (7.5)$$

Comparando as duas fórmulas em (7.5), vemos que a única diferença é o uso do fator de correção para populações finitas $\sqrt{(N-n)/(N-1)}$.

Como ocorreu com a média amostral $\bar{x}$, a diferença entre as expressões relativas à população finita e à população infinita torna-se desprezível se o tamanho da população for grande em comparação com o tamanho da amostra. Seguimos a mesma regra prática que recomendamos em relação à média amostral. Ou seja, se a população for finita, com $n/N \leq 0{,}05$, usaremos $\sigma_{\bar{p}} = \sqrt{p(1-p)/n}$. Entretanto, se a população for finita, com $n/N > 0{,}05$, o fator de correção para populações finitas deverá ser usado. Novamente, a menos que seja especificamente indicado, ao longo de todo o livro presumiremos que o tamanho da população seja grande em relação ao tamanho da amostra e, desse modo, o fator de correção para populações finitas é desnecessário.

Na Seção 7.5, utilizamos a expressão erro padrão da média para nos referir ao desvio padrão de $\bar{x}$. Afirmamos que, em geral, o termo *erro padrão* refere-se ao desvio padrão de um estimador por ponto. Dessa forma, quanto às proporções, utilizamos a expressão *erro padrão da proporção* para nos referir ao desvio padrão de $\bar{p}$. Retornemos agora ao Exemplo da EAI e calculemos o erro padrão da proporção associada às amostras aleatórias simples de seus 30 gerentes.

Em relação ao estudo da EAI, sabemos que a proporção da população de gerentes que participaram do programa de treinamento gerencial é $p = 0{,}60$. Com $n/N = 30/2.500 = 0{,}012$, podemos ignorar o fator de correção para populações finitas quando calculamos o desvio padrão da proporção. Para a amostra aleatória simples de 30 gerentes, $\sigma_{\bar{p}}$ é

$$\sigma_{\bar{p}} = \sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{0{,}60(1-0{,}60)}{30}} = 0{,}0894$$

Forma da Distribuição Amostral de $\bar{p}$

Agora que conhecemos a média e o desvio padrão da distribuição amostral de $\bar{p}$, a etapa final consiste em determinarmos o formato, ou forma, da distribuição amostral. A proporção amostral é $\bar{p} = x/n$. Para uma amostra aleatória simples de uma população grande, o valor de $\bar{p}$ é uma variável aleatória binomial que indica o número de elementos contidos na amostra que possuem a característica de interesse. Uma vez que n é uma constante, a probabilidade de x/n é idêntica à probabilidade binomial de x, o que significa que a distribuição amostral de $\bar{p}$ também é uma distribuição discreta de probabilidade e que a probabilidade correspondente a cada valor de x/n é idêntica à probabilidade binomial de x.

No Capítulo 6, também mostramos que uma distribuição binomial pode ser aproximada por meio de uma distribuição normal sempre que o tamanho da amostra for grande o bastante para satisfazer às duas condições seguintes:

$$np \geq 5 \text{ e } n(1-p) \geq 5$$

Considerando que essas duas condições tenham sido satisfeitas, a distribuição de probabilidade de x, que é o número de elementos na amostra que possuem a característica de interesse, pode ser aproximada por meio de uma distribuição normal. E, desde que n seja uma constante, a distribuição amostral de $\bar{p} = x/n$ também pode ser aproximada por meio de uma distribuição normal. Essa aproximação é definida da seguinte maneira:

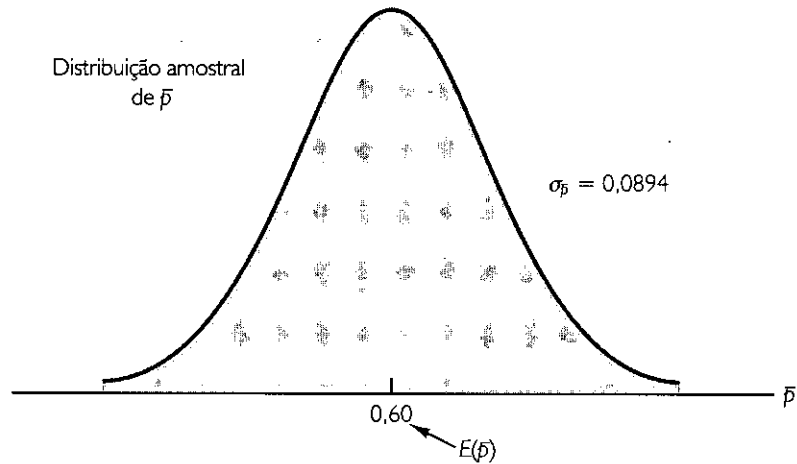
A distribuição amostral de $\bar{p}$ pode ser aproximada por meio de uma distribuição normal sempre que $np \geq 5$ e $n(1-p) \geq 5$.

Em aplicações práticas, quando desejamos uma estimativa da proporção de uma população, descobrimos que os tamanhos das amostras quase sempre são suficientemente grandes para permitir o uso de uma aproximação normal à distribuição amostral de $\bar{p}$.

Lembre-se de que, em relação ao problema de amostragem da EAI, sabemos que a proporção da população de gerentes que participaram do programa de treinamento é $p = 0{,}60$. Com uma amostra aleatória simples de tamanho 30, temos $np = 30(0{,}60) = 18$, e $n(1-p) = 30(0{,}40) = 12$. Então, a distribuição amostral de $\bar{p}$ pode ser aproximada pela distribuição normal apresentada na Figura 7.8.

Valor Prático da Distribuição Amostral de $\bar{p}$

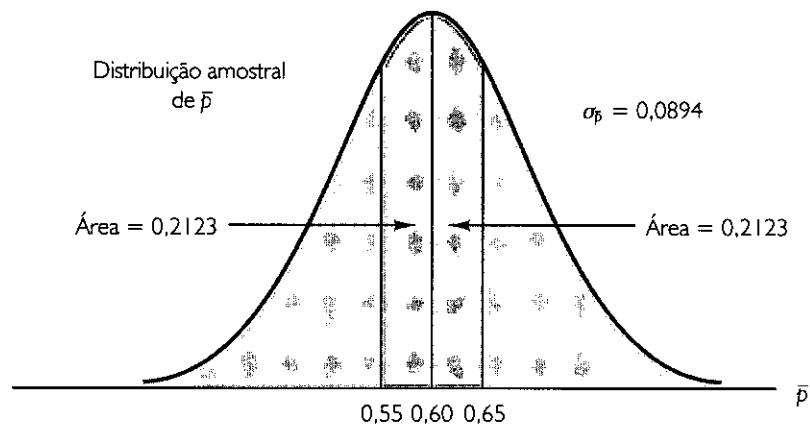
O valor prático da distribuição amostral $\bar{p}$ é que ela pode ser usada para produzir informações probabilísticas a respeito da diferença entre a proporção amostral e a proporção populacional. Por exemplo, suponha que no problema da EAI o diretor de pessoal queira saber qual é a probabilidade de obter um valor de $\bar{p}$ que se situe no intervalo de 0,05 da proporção populacional de gerentes da EAI que participaram do programa de treinamento. Ou seja, qual é a probabilidade de obter uma amostra com uma proporção amostral $\bar{p}$ que se situe entre 0,55 e 0,65? A área com sombreamento mais escuro na Figura 7.9 representa essa probabilidade. Usando o fato de que a distribuição amostral de $\bar{p}$ pode ser aproximada por uma distribuição normal com uma média igual a 0,60 e desvio padrão da proporção igual a $\sigma_{\bar{p}} = 0{,}894$, descobrimos que a variável aleatória normal padrão correspondente a $\bar{p} = 0{,}55$ tem o valor $z = (0{,}55 - 0{,}60)/0{,}0894 = -0{,}56$. Consultando a tabela de distribuição normal padrão, notamos que a área entre $z = -0{,}56$ e $z = 0$ é 0,2123.

Figura 7.8 Distribuição amostral de $\bar{p}$ referente à proporção de gerentes da EAI que participaram do programa de treinamento gerencial

Similarmente, para $\bar{p}$= 0,65, descobrimos que a área entre $z = 0$ e $z = 0,56$ é 0,2123. Desse modo, a probabilidade de selecionar uma amostra que forneça uma proporção amostral $\bar{p}$ dentro de 0,05 da proporção populacional p é 0,2123 + 0,2123 = 0,4246.

Se pensarmos em aumentar o tamanho da amostra para $n = 100$, o erro padrão da proporção se torna

$$\sigma_{\bar{p}} = \sqrt{\frac{0,60(1 - 0,60)}{100}} = 0,049$$

Com um tamanho de amostra de 100 gerentes da EAI, a probabilidade de a proporção amostral ter um valor dentro de 0,05 da proporção populacional agora pode ser calculada. Uma vez que a distribuição amostral é aproximadamente normal, com média de 0,60 e erro padrão igual a 0,49, podemos usar a tabela de distribuição normal padrão para encontrar a área, ou probabilidade. Para $\bar{p} = 0,55$, temos $z = (0,55 - 0,60)/0,49 = -1,02$. Consultando a tabela de distribuição normal padrão, vemos que a área entre $z = -1,02$ e $z = 0$ é 0,3461. De forma semelhante, para 0,65, a área entre $z = 0$ e $z = 1,02$ é 0,3461. Assim, se o tamanho da amostra for aumentado de 30 para 100, a probabilidade de a proporção amostral $\bar{p}$ estar dentro de 0,05 da proporção populacional p se elevará para 0,3461 + 0,3461 = 0,6922.

Figura 7.9 Probabilidade de se obter $\bar{p}$ entre 0,55 e 0,65

Exercícios

Métodos

31. Uma amostra aleatória simples de tamanho 100 é selecionada de uma população com $p = 0,40$.
   a. Qual é o valor esperado de $\bar{p}$?
   b. Qual é o erro padrão de $\bar{p}$?
   c. Apresente a distribuição amostral de $\bar{p}$.
   d. O que a distribuição amostral de $\bar{p}$ indica?

AUTOTESTE

32. A proporção de uma população é 0,40. Uma amostra aleatória simples de tamanho 200 será tomada e a proporção amostral $\bar{p}$ será usada para estimar a proporção da população.
   a. Qual é a probabilidade de a proporção amostral estar dentro de ±0,03 da proporção populacional?
   b. Qual é a probabilidade de a proporção amostral estar dentro de ±0,05 da proporção populacional?
33. Suponha que a proporção populacional seja 0,55. Calcule o erro padrão da proporção, $\sigma_{\bar{p}}$, para os tamanhos de amostra 100, 200, 500 e mil. O que se pode dizer sobre o tamanho do erro padrão da proporção quando o tamanho da amostra é aumentado?
34. A proporção populacional é 0,30. Qual é a probabilidade de a proporção amostral estar dentro de ±0,04 da proporção populacional correspondente a cada um dos seguintes tamanhos de amostra?
   a. $n = 100$
   b. $n = 200$
   c. $n = 500$
   d. $n = 1.000$
   e. Qual é a vantagem de um tamanho de amostra maior?

Aplicações

AUTOTESTE

35. O presidente da Doerman Distributors, Inc., acredita que 30% das encomendas feitas à firma são provenientes de clientes que compram pela primeira vez. Uma amostra aleatória simples de 100 pedidos será usada para estimar a proporção de clientes que compram pela primeira vez.
   a. Suponha que o presidente esteja correto e $p = 30$. Qual é a distribuição amostral de $\bar{p}$ nesse estudo?
   b. Qual é a probabilidade de a proporção amostral $\bar{p}$ estar entre 0,20 e 0,40?
   c. Qual é a probabilidade de a proporção amostral estar entre 0,25 e 0,35?
36. A *Business Week* divulgou que 56% das famílias dos Estados Unidos têm acesso à internet (*Business Week*, 21 de maio de 2001). Use a proporção populacional $p = 0,56$ e suponha que uma amostra de 300 famílias seja selecionada.
   a. Apresente a distribuição amostral de $\bar{p}$, em que $\bar{p}$ é a proporção amostral de famílias que têm acesso à internet.
   b. Qual é a probabilidade de a proporção amostral estar dentro de ±0,03 da proporção populacional?
   c. Responda ao item (b) considerando os tamanhos de amostra 600 e mil.
37. Pesquisas eleitorais da *Time*/CNN monitoraram a opinião pública em relação aos candidatos presidenciais durante a campanha eleitoral à Presidência da República de 2000. Uma pesquisa patrocinada pela *Time*/CNN e realizada pela Yankelovich Partners, Inc., usou uma amostra de 589 eleitores (*Time*, 26 de junho de 2000). Suponha que a proporção populacional correspondente a um candidato presidencial seja $p = 0,50$. Admitamos que $\bar{p}$ seja a proporção amostral de eleitores provavelmente favoráveis ao candidato presidencial.
   a. Apresente a distribuição amostral de $\bar{p}$.
   b. Qual é a probabilidade de a pesquisa da *Time*/CNN produzir uma proporção amostral dentro de ±0,04 da proporção populacional?
   c. Qual é a probabilidade de a pesquisa da *Time*/CNN produzir uma proporção amostral dentro de ±0,03 da proporção populacional?
   d. Qual é a probabilidade de a pesquisa da *Time*/CNN produzir uma proporção amostral dentro de ±0,02 da proporção populacional?
38. A Roper ASW promoveu uma pesquisa para saber qual era a postura dos norte-americanos adultos em relação a dinheiro e felicidade (*Money*, outubro de 2003). Cinqüenta e seis por cento dos entrevistados disseram que faziam um balanço de seus talões de cheque pelo menos uma vez por mês.

a. Suponha que uma amostra de 400 norte-americanos adultos tenha sido tomada. Apresente a distribuição amostral da proporção de adultos que controlam seus talões de cheque pelo menos uma vez por mês.
b. Qual é a probabilidade de a proporção amostral estar dentro de ±0,02 da proporção populacional?
c. Qual é a probabilidade de a proporção amostral estar dentro de ±0,04 da proporção populacional?

39. O *Democrat and Chronicle* divulgou que 25% dos vôos que chegaram ao aeroporto de San Diego durante os cinco primeiros meses de 2001 estavam atrasados (*Democrat and Chronicle*, 23 de julho de 2001). Suponha que a proporção populacional seja $p = 0,25$.
a. Apresente a distribuição amostral de $\bar{p}$, a proporção de vôos atrasados em uma amostra de mil vôos.
b. Qual é a probabilidade de a proporção amostral estar dentro de ±0,03 da proporção populacional se uma amostra de tamanho mil for selecionada?
c. Responda ao item (b) considerando uma amostra de 500 vôos.

40. A Grocery Manufacturers of America divulgou que 76% dos consumidores lêem os ingredientes relacionados no rótulo dos produtos. Suponha que a proporção populacional seja $p = 0,76$ e que uma amostra de 400 consumidores seja selecionada da população.
a. Apresente a distribuição amostral da proporção da amostra $\bar{p}$, em que $\bar{p}$ é a proporção dos consumidores amostrados que lêem os ingredientes relacionados no rótulo do produto.
b. Qual é a probabilidade de a proporção amostral estar dentro de ±0,03 da proporção populacional?
c. Responda ao item (b) considerando uma amostra de 750 consumidores.

41. O Food Marketing Institute indica que 17% das famílias gastam mais de US$ 100,00 por semana em produtos de mercearia. Suponha que a proporção populacional seja $p = 0,17$ e que uma amostra aleatória simples de 800 famílias seja selecionada da população.
a. Apresente a distribuição amostral de $\bar{p}$, que é a proporção amostral de famílias que gastam mais do que US$ 100,00 por semana em produtos de mercearia.
b. Qual é a probabilidade de a proporção amostral estar dentro de ±0,02 da proporção populacional?
c. Responda ao item (b) considerando uma amostra de 1.600 famílias.

7.7 MÉTODOS DE AMOSTRAGEM

Esta seção apresenta uma breve introdução a outros métodos de amostragem, diferentes da amostragem aleatória simples.

Descrevemos o procedimento de amostragem aleatória simples e discutimos as propriedades das distribuições amostrais de $\bar{x}$ e $\bar{p}$ quando se usa a amostragem aleatória simples. Entretanto, a amostragem aleatória simples não é o único método de amostragem disponível. Métodos como a amostragem aleatória estratificada, a amostragem por conglomerados e a amostragem sistemática apresentam vantagens sobre a amostragem aleatória simples em algumas situações. Nesta seção, apresentaremos brevemente esses métodos alternativos de amostragem.

Amostragem Aleatória Estratificada

Na **amostragem aleatória estratificada**, os elementos da população são divididos primeiramente em grupos denominados *estratos*, de forma que cada elemento da população pertença a um e somente a um estrato. A base para formação dos estratos, por exemplo, departamento, local, idade, tipo de indústria etc. ficam a critério do projetista da amostra. Porém, os melhores resultados são obtidos quando os elementos contidos em cada estrato são o mais similares possível. A Figura 7.10 representa o diagrama de uma população dividida em H estratos.

Figura 7.10 Diagrama da amostragem aleatória estratificada

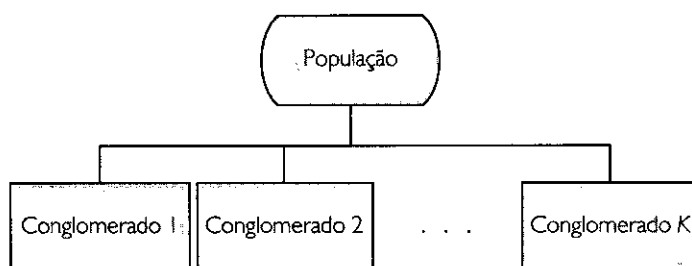
Depois que os estratos são formados, extrai-se uma amostra aleatória simples de cada um deles. Há fórmulas disponíveis para se combinar os resultados das amostras de estrato individuais em uma estimativa do parâmetro populacional de interesse. O valor da amostragem aleatória estratificada depende da homogeneidade dos elementos contidos nos estratos. Se os elementos contidos nos estratos forem similares, os estratos terão baixas variâncias. Desse modo, tamanhos de amostra relativamente pequenos podem ser usados para se obter boas estimativas das características dos estratos. Se os estratos forem homogêneos, o procedimento de amostragem aleatória estratificada produzirá resultados tão precisos quanto os da amostragem aleatória simples, mas utilizando um tamanho total de amostra menor.

A amostragem aleatória estratificada funciona melhor quando a variância entre os elementos de cada estrato é relativamente pequena.

Amostragem por Conglomerados

Na **amostragem por conglomerados**, os elementos da população são divididos primeiramente em grupos distintos denominados *conglomerados*. Cada elemento da população pertence a um e somente a um conglomerado (veja a Figura 7.11). Extrai-se, então, uma amostra aleatória simples dos conglomerados. Todos os elementos contidos em cada conglomerado amostrado formam a amostra. A amostragem por conglomerados tende a produzir os melhores resultados quando os elementos neles contidos não são similares. No caso ideal, cada conglomerado é uma versão representativa em pequena escala da população inteira. O valor da amostragem por conglomerados depende de quão representativo é cada conglomerado da população inteira. Se todos os conglomerados forem similares nesse sentido, a amostragem de um pequeno número de conglomerados produzirá boas estimativas dos parâmetros populacionais.

A amostragem por conglomerados funciona melhor quando cada agrupamento produz uma representação em pequena escala da população.

Figura 7.11 Diagrama da amostragem por conglomerados

Uma das principais aplicações da amostragem por conglomerados é a amostragem por áreas, em que os conglomerados são quarteirões de uma cidade ou outras áreas bem definidas. A amostragem por conglomerados geralmente requer um tamanho maior de amostra total do que a amostragem aleatória simples ou a amostragem aleatória estratificada. Entretanto, ela pode resultar em economias de custo pelo fato de que, quando um entrevistador é enviado a um conglomerado amostrado (por exemplo, uma localização em um quarteirão da cidade), muitas observações amostrais podem ser obtidas em um tempo relativamente breve. Portanto, um tamanho maior de amostra pode ser obtido com um custo total significativamente menor.

Amostragem Sistemática

Em algumas situações de amostragem, especialmente quando se trata de grandes populações, consome muito tempo selecionar uma amostra aleatória simples pelo método de encontrar primeiramente um número aleatório e depois contar ou pesquisar a lista da população até que o elemento correspondente seja encontrado. Uma alternativa à amostragem aleatória simples é a **amostragem sistemática**. Por exemplo, quando se deseja um tamanho de amostra 50 de uma população que contém 5 mil elementos, extrai-se como amostra um elemento em cada 5.000/50 = 100 elementos da população. Uma amostra sistemática, nesse caso, envolve selecionar aleatoriamente um dos primeiros 100 elementos da lista da população. Os outros elementos da amostra são identificados começando-se com o primeiro elemento amostrado e selecionando-se então cada 100º elemento seguinte na lista da população. Com efeito, a amostra de 50 é identificada deslocando-se sistematicamente entre a população e identificando-se cada 100º elemento seguinte ao primeiro elemento selecionado aleatoriamente. A amostra de 50 geralmente será mais fácil de identificar dessa maneira do que se usássemos a amostragem aleatória simples. Como o primeiro elemento selecionado é uma escolha

aleatória, geralmente se presume que uma amostra sistemática tem as propriedades de uma amostra aleatória simples. Essa hipótese é aplicável especialmente quando a lista de elementos da população apresenta uma organização aleatória dos elementos.

Amostragem de Conveniência

Os métodos de amostragem discutidos até agora são chamados técnicas de *amostragem probabilística*. Os elementos selecionados da população têm uma probabilidade conhecida de serem incluídos na amostra. A vantagem da amostragem probabilística é que a distribuição amostral apropriada da estatística da amostra geralmente pode ser identificada. Fórmulas como as da amostragem aleatória simples, apresentadas neste capítulo, podem ser utilizadas para determinarmos as propriedades da distribuição amostral. Depois, a distribuição amostral pode ser usada para fazermos afirmações probabilísticas a respeito do erro associado aos resultados amostrais.

A **amostragem de conveniência** é uma técnica de *amostragem não-probabilística*. Como o nome implica, a amostra é identificada primeiramente por conveniência. Elementos são incluídos na amostra sem probabilidades previamente especificadas ou conhecidas de eles serem selecionados. Por exemplo, um professor que faz pesquisas em uma universidade pode utilizar estudantes voluntários para compor uma amostra, simplesmente porque eles estão disponíveis e participarão como objetos de experiência por pouco ou nenhum custo. Analogamente, um inspetor pode extrair uma amostra de um embarque de laranjas selecionando-as casualmente de vários engradados. Rotular cada laranja e usar o método probabilístico de amostragem seria impraticável. Amostras tais como de animais selvagens capturados e de grupos de voluntários para pesquisa de consumidores também são amostras de conveniência.

As amostras de conveniência têm a vantagem de permitir que a escolha de amostras e a coleta de dados sejam relativamente fáceis; entretanto, é impossível avaliar a "excelência" da amostra em termos de sua representatividade da população. Uma amostra de conveniência tanto pode produzir bons resultados como não; nenhum procedimento estatisticamente justificável possibilita uma análise de probabilidade e inferência sobre a qualidade dos resultados da amostra.

Às vezes, os pesquisadores aplicam a amostras de conveniência certos métodos estatísticos projetados especificamente para amostras probabilísticas, argumentando que uma amostra de conveniência pode ser tratada como se fosse uma amostra probabilística. Entretanto, esse argumento não é sustentável, e devemos ser cautelosos ao interpretar os resultados das amostras de conveniência que são utilizados para fazer inferências sobre populações.

Amostragem de Julgamento

Uma técnica adicional de amostragem não-probabilística é a **amostragem de julgamento**. Nessa abordagem, a pessoa que conhece mais profundamente o tema do estudo escolhe os elementos que julga serem os mais representativos da população. Freqüentemente, esse método é uma maneira relativamente fácil de selecionar uma amostra. Por exemplo, um repórter pode tomar como amostra dois ou três senadores, julgando que eles refletem a opinião geral de todos os senadores. Entretanto, a qualidade dos resultados da amostra depende do julgamento da pessoa que a seleciona. Novamente, recomendamos muita cautela ao tirar conclusões baseadas em amostras de julgamento que são utilizadas para fazer inferências sobre populações.

NOTAS E COMENTÁRIOS

Recomendamos o uso de métodos de amostragem probabilística: amostragem aleatória simples, amostragem aleatória estratificada, amostragem por conglomerados ou amostragem sistemática. Em relação a esses métodos, há fórmulas disponíveis para avaliar a "excelência" dos resultados amostrais em termos de quão próximos eles estão dos parâmetros populacionais a serem determinados. Uma avaliação da excelência não pode ser feita com base em amostragens de conveniência ou de julgamento. Desse modo, devemos tomar muito cuidado ao interpretar resultados baseados em métodos de amostragem não-probabilísticos.

Resumo

Neste capítulo apresentamos os conceitos de amostragem aleatória simples e de distribuições amostrais. Demonstramos como uma amostra aleatória simples pode ser selecionada e como os dados coletados para a amostra podem ser utilizados para desenvolvermos estimações por ponto dos parâmetros populacionais. Uma vez que diferentes amostras aleatórias simples produzem diferentes valores para os estimadores por ponto, os estimadores por ponto como $\bar{x}$ e $\bar{p}$ são variáveis aleatórias. A distribuição probabilística desse tipo de variável aleatória denomina-se distribuição amostral. Em especial, descrevemos as distribuições amostrais da média amostral $\bar{x}$ e da proporção amostral $\bar{p}$.

Ao considerar as características das distribuições amostrais de $\bar{x}$ e $\bar{p}$, estabelecemos que $E(\bar{x}) = \mu$ e $E(\bar{p}) = p$. Depois de desenvolvermos as fórmulas do desvio padrão, ou erro padrão, desses estimadores, descrevemos as condições necessárias para que as distribuições amostrais de $\bar{x}$ e $\bar{p}$ sigam uma distribuição normal. Foram discutidos outros métodos de amostragem, entre os quais se contam a amostragem aleatória estratificada, a amostragem por conglomerados, a amostragem sistemática, a amostragem de conveniência e a amostragem de julgamento.

Glossário

Parâmetro Uma característica numérica da população, como a média populacional μ, o desvio padrão da população σ, a proporção populacional p e assim por diante.

Amostragem aleatória simples População finita: uma amostragem escolhida de maneira que cada amostra possível de tamanho n tenha a mesma probabilidade de ser selecionada. População infinita: uma amostra selecionada de tal forma que cada elemento vem da mesma população e os elementos são selecionados independentemente.

Amostragem sem substituição Tão logo um elemento é incluído na amostra, ele é eliminado da população e não pode ser escolhido uma segunda vez.

Amostragem com substituição Tão logo um elemento é incluído na amostra, ele é devolvido à população. Um elemento selecionado anteriormente pode ser selecionado novamente e, portanto, pode aparecer na amostra mais de uma vez.

Estatística da amostra Uma característica da amostra, por exemplo, uma média amostral $\bar{x}$, um desvio padrão da amostra s, uma proporção da amostra $\bar{p}$ e assim por diante. O valor da estatística da amostra é usado para estimar o valor do parâmetro populacional correspondente.

Estimação por ponto A estatística da amostra, por exemplo, $\bar{x}$, s ou $\bar{p}$, que fornece o estimador por ponto do parâmetro populacional.

Estimativa por ponto O valor de um estimador por ponto usado em um caso em particular como estimativa de um parâmetro populacional.

Distribuição amostral Uma distribuição de probabilidade que consiste em todos os valores possíveis de uma estatística amostral.

Sem viés Uma propriedade de um estimador por ponto que está presente quando o valor esperado do estimador por ponto é igual ao parâmetro populacional que ele estima.

Fator de correção para populações finitas O termo $\sqrt{(N - n)/(N - 1)}$ que é usado nas fórmulas para $\sigma_{\bar{x}}$ e $\sigma_{\bar{p}}$, sempre que uma população finita, em vez de uma infinita, é amostrada. A regra prática geralmente aceita é ignorar o fator de correção para populações finitas sempre que $n/N \leq 0{,}05$.

Erro padrão O desvio padrão de um estimador por ponto.

Teorema do limite central Um teorema que nos possibilita usar a distribuição normal de probabilidade para fazer a aproximação à distribuição amostral de $\bar{x}$ sempre que o tamanho da amostra for grande.

Amostragem aleatória estratificada Um método de amostragem probabilística no qual a população primeiramente é dividida em estratos e então se toma uma amostra aleatória simples de cada estrato.

Amostragem por conglomerados Um método de amostragem probabilística no qual a população primeiramente é dividida em conglomerados e então se toma uma amostra aleatória simples dos aglomerados.

Amostragem sistemática Um método de amostragem probabilística no qual selecionamos aleatoriamente um dos primeiros k elementos e depois selecionamos cada k-ésimo elemento seguinte.

Amostragem de conveniência Um método não-probabilístico de amostragem em que os elementos são selecionados para a amostra com base na conveniência.

Amostragem de julgamento Um método não-probabilístico de amostragem em que os elementos são selecionados para a amostra com base no julgamento da pessoa que realiza o estudo.

Fórmulas-Chave

Valor Esperado de $\bar{x}$

$$E(\bar{x}) = \mu \quad (7.1)$$

Desvio Padrão (Erro padrão) de $\bar{x}$

População Finita — *População Infinita*

$$\sigma_{\bar{x}} = \sqrt{\frac{N-n}{N-1}}\left(\frac{\sigma}{\sqrt{n}}\right) \qquad \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \quad (7.2)$$

Valor Esperado de $\bar{p}$

$$E(\bar{p}) = p \quad (7.4)$$

Desvio Padrão (Erro padrão) de $\bar{p}$

População Finita — *População Infinita*

$$\sigma_{\bar{p}} = \sqrt{\frac{N-n}{N-1}}\sqrt{\frac{p(1-p)}{n}} \qquad \sigma_{\bar{p}} = \sqrt{\frac{p(1-p)}{n}} \quad (7.5)$$

Exercícios Suplementares

42. O Corporate Scoreboard (Placar Corporativo) da *Business Week* publica dados trimestrais sobre vendas, lucros, renda líquida, retorno sobre o patrimônio líquido, razão preço/rendimentos, e rendimentos por ação de 899 empresas (*Business Week*, 14 de agosto de 2000). As empresas podem ser enumeradas de 1 a 899 na ordem em que aparecem na lista do Corporate Scoreboard. Inicie na parte inferior da segunda coluna de dígitos aleatórios da Tabela 7.1. Ignorando os dois primeiros dígitos de cada grupo e usando números aleatórios de três dígitos que se iniciam com 12, leia a coluna de baixo para *cima* para identificar o número (de 1 a 899) das oito primeiras empresas a serem incluídas em uma amostra aleatória simples.

43. O povo norte-americano demonstra uma preocupação cada vez maior com os crescentes custos dos planos de saúde. Em 1990, a média de gastos anuais por segurado era US$ 3.267; em 2003, a média de gastos anuais por segurado era US$ 6.883 (*Money*, outono de 2003). Suponha que você contratou uma empresa de consultoria para tomar uma amostra de 50 segurados em 2003 para aprofundar a investigação sobre a natureza dos gastos. Suponha que o desvio padrão da população para 2003 tenha sido US$ 2 mil.
    a. Apresente a distribuição amostral da quantia média de gastos com planos de saúde correspondente a uma amostra de 50 segurados em 2003.
    b. Qual é a probabilidade de a média amostral estar dentro de ±US$ 300 da média da população?
    c. Qual é a probabilidade de a média amostral ser maior do que US$ 7.500? Se a empresa de consultoria lhe disser que a média amostral dos segurados que entrevistaram foi de US$ 7.500, você perguntaria se eles seguiram procedimentos corretos de amostragem aleatória simples? Por quê?

44. A *Business Week* pesquisou ex-alunos de cursos de MBA dez anos após a graduação (*Business Week*, 22 de setembro de 2003). Uma revelação foi que os ex-alunos gastam em média US$ 115,50 por semana com almoços ou jantares sociais. Você foi solicitado a realizar um estudo de acompanhamento, tomando uma amostra de 40 desses ex-alunos de MBA. Suponha que o desvio médio da população seja US$ 35,00.
    a. Apresente a distribuição amostral de $\bar{x}$, a média amostral de gastos semanais dos 40 ex-alunos de MBA.
    b. Qual é a probabilidade de a média amostral estar dentro de US$ 10 da média da população?
    c. Suponha que você encontre uma média amostral de US$ 100. Qual é a probabilidade de encontrar uma média amostral de US$ 100 ou menos? Você consideraria essa amostra é de um grupo de ex-alunos com gastos incomumente baixos? Por quê?

45. A média de tempo que os norte-americanos passam assistindo à televisão é de 15 horas por semana (*Money*, novembro de 2003). Suponha que uma amostra de 60 norte-americanos seja tomada para que se investigue com mais profundidade os hábitos relativos à TV. Suponha que o desvio padrão da população referente ao tempo semanal que passam assistindo à TV seja $\sigma = 4$ horas.

a. Qual é a probabilidade de a média da amostra estar dentro de 1 hora da média da população?
b. Qual é a probabilidade de a média da amostra estar dentro de 45 minutos da média da população?

46. O salário anual médio dos servidores públicos federais do estado de Indiana é US$ 41.979 (*The World Almanac 2001*). Use esse valor como média populacional e suponha que o desvio padrão da população seja σ = US$ 5.000. Suponha que uma amostra aleatória de 50 servidores públicos federais seja selecionada da população.

a. Qual é o valor do erro padrão da média?
b. Qual é a probabilidade de a média da amostra ser maior que US$ 41.979?
c. Qual é a probabilidade de a média da amostra estar dentro de US$ 1.000 da média da população?
d. Como a probabilidade do item (c) se alteraria se o tamanho da amostra fosse aumentado para 100?

47. Três firmas têm inventários que diferem quanto ao tamanho. O inventário da firma A contém 2 mil itens, o inventário da firma B contém 5 mil itens, e o inventário da firma C contém 10 mil itens. O desvio padrão da população quanto ao custo dos itens é $\sigma = 144$. Um consultor em estatística recomenda que cada firma extraia uma amostra de 50 itens de seu inventário para produzir estimativas estatisticamente válidas do custo médio por item. Os gerentes da pequena empresa declaram que, já que ela possui a menor população, seriam capazes de fazer a estimativa utilizando uma amostra muito menor do que seria necessária para empresas maiores. Entretanto, o consultor afirma que para obter o mesmo desvio padrão e, desse modo, a mesma precisão nos resultados amostrais, todas as empresas devem usar o mesmo tamanho de amostra, independentemente do tamanho da população.

a. Usando o fator de correção para populações finitas, calcule o erro padrão correspondente a cada uma das três firmas, dada uma amostra de tamanho 50.
b. Qual é a probabilidade de a média amostral $\bar{x}$ correspondente a cada uma das firmas estar dentro de ± 25 da média populacional μ?

48. Um pesquisador relata os resultados de uma pesquisa afirmando que o erro padrão da média é 20. O desvio padrão da população é 500.

a. Qual é o tamanho da amostra utilizada nessa pesquisa?
b. Qual é a probabilidade de a estimação por ponto estar dentro de ± 25 da média da população?

49. Um processo de produção é checado periodicamente por um inspetor de controle da qualidade. O inspetor seleciona amostras aleatórias simples de 30 produtos acabados e calcula a média $\bar{x}$ de peso dos produtos da amostra. Se os resultados dos testes realizados no decorrer de um longo período mostram que 5% dos valores de $\bar{x}$ estão acima de 2,1 libras (0,95 kg) e que 5% estão abaixo de 1,9 libra (0,86 kg), quais são a média e o desvio padrão da população de produtos produzidos sob esse processo?

50. Em 13 de junho de 2001, 30,5% dos investidores individuais eram altistas (*bullish*) no mercado de títulos de curto prazo (*AAII Journal*, julho de 2001). Responda às seguintes questões considerando que seja usada uma amostra de 200 investidores individuais.

a. Apresente a distribuição amostral de $\bar{p}$, a proporção amostral de investidores individuais que são altistas no mercado de títulos de curto prazo.
b. Qual é a probabilidade de a proporção amostral estar dentro de $\pm 0{,}04$ da proporção populacional?
c. Qual é a probabilidade de a proporção amostral estar dentro de $\pm 0{,}02$ da proporção populacional?

51. Uma firma de pesquisa de mercado realiza pesquisas telefônicas com um índice histórico de 40% de respostas. Qual é a probabilidade de, em uma nova amostra de 400 números telefônicos, pelo menos 150 pessoas colaborarem e responderem às perguntas? Em outras palavras, qual é a probabilidade de a proporção da amostra ser de, pelo menos, $150/400 = 0{,}375$?

52. De acordo com a ORC International, 71% dos usuários da internet conectam seus computadores à rede por meio de linhas telefônicas convencionais (*USA Today*, 18 de janeiro de 2000). Considere uma proporção populacional $p = 0{,}71$.

a. Qual é a probabilidade de a proporção amostral de uma amostra aleatória simples de 350 usuários da internet estar dentro de $\pm 0{,}05$ da proporção populacional?
b. Qual é a probabilidade de a proporção amostral de uma amostra aleatória simples de 350 usuários da internet ser de 0,75 ou maior?

53. A proporção de indivíduos segurados pela All-Driver Automobile Insurance Company que receberam pelo menos uma multa de trânsito durante um período de cinco anos é 0,15.

a. Mostre qual seria a distribuição amostral de $\bar{p}$ se uma amostra aleatória de 150 indivíduos segurados fosse utilizada para estimar a proporção dos que receberam pelo menos uma multa de trânsito.

b. Qual é a probabilidade de a proporção da amostra estar dentro de ±0,03 da proporção da população?

54. Lori Jeffrey é uma bem-sucedida representante de vendas de uma grande editora de livros universitários. Historicamente, Lori consegue fazer que adotem um livro em 25% de seus contatos de vendas. Tomando seus contatos de vendas realizados durante um mês como uma amostra de todos os contatos de vendas possíveis, considere que uma análise estatística dos dados produza um erro padrão da proporção igual a 0,0625.

a. Qual é o tamanho da amostra usada nessa análise? Ou seja, quantos contatos de vendas Lori fez durante o mês?

b. Considere que $\bar{p}$ indica a proporção amostral de adoção de livros durante o mês. Apresente a distribuição amostral $\bar{p}$.

c. Usando a distribuição amostral de $\bar{p}$, calcule a probabilidade de Lori fazer que adotem livros em 30% ou mais de seus contatos de vendas durante o período de um mês?

Apêndice 7.1 – Amostragem Aleatória com o Minitab

Se uma relação dos elementos de uma população estiver disponível em um arquivo do Minitab, esse programa pode ser usado para selecionar uma amostra aleatória simples. Por exemplo, uma relação das 100 maiores regiões metropolitanas dos Estados Unidos e do Canadá é apresentada na coluna 1 do conjunto de dados (*data set*) MetAreas (*Places Rated Almanac – The Millennium Edition 2000*). A coluna 2 contém uma classificação global de cada região metropolitana. As dez primeiras regiões metropolitanas do conjunto de dados e suas classificações correspondentes são apresentadas na Tabela 7.6.

Suponha que você queira selecionar uma amostra aleatória simples de 30 regiões metropolitanas a fim de realizar um estudo detalhado do custo de vida nos Estados Unidos e no Canadá. As etapas a seguir podem ser usadas para selecionar a amostra.

Etapa 1. Selecione o menu **Calc**
Etapa 2. Escolha **Random Data**
Etapa 3. Escolha **Sample From Columns**
Etapa 4. Quando a caixa de diálogo Sample From Columns aparecer:
Digite 30 na caixa **Sample**
Digite C1 C2 na caixa de baixo
Digite C3 C4 na caixa **Store samples in**
Etapa 5. Dê um clique em **OK**

A amostra aleatória de 30 regiões metropolitanas aparecerá nas colunas C3 e C4.

Apêndice 7.2 – Amostragem Aleatória com o Excel

Se uma relação dos elementos de uma população estiver disponível em um arquivo do Excel, esse programa pode ser usado para selecionar uma amostra aleatória simples. Por exemplo, uma relação das 100 maiores regiões metropolitanas dos Estados Unidos e do Canadá é apresentada na coluna A do conjunto de dados (*data set*) MetAreas (*Places Rated Almanac – The Millennium Edition 2000*). A coluna B contém uma classificação global de cada região metropolitana. As dez primeiras regiões metropolitanas do conjunto de dados e suas classificações correspondentes são apresentadas na Tabela 7.6. Suponha que você queira selecionar uma amostra aleatória simples de 30 regiões metropolitanas a fim de realizar um estudo detalhado do custo de vida nos Estados Unidos e no Canadá.

Tabela 7.6 Classificação global das dez primeiras regiões metropolitanas do conjunto de dados (*data set*) MetAreas

Região Metropolitana	Classificação	Região Metropolitana	Classificação
Albany, NY	64,18	Baltimore, MD	69,75
Albuquerque, NM	66,16	Birmingham, AL	69,59
Appleton, WI	60,56	Boise City, ID	68,36
Atlanta, GA	69,97	Boston, MA	68,99
Austin, TX	71,48	Buffalo, NY	66,10

ARQUIVO DA INTERNET
MetAreas

As linhas de qualquer conjunto de dados do Excel podem ser dispostas em ordem aleatória acrescentando-se uma coluna extra ao conjunto de dados e preenchendo-se a coluna com números aleatórios com o uso da função =ALEATÓRIO ( ). Então, usando-se a capacidade de classificação em ordem crescente do Excel na coluna de números aleatórios, as linhas do conjunto de dados serão reorganizadas aleatoriamente. A amostra aleatória de tamanho n aparecerá nas primeiras n linhas do conjunto de dados reorganizado.

No conjunto de dados MetAreas, os rótulos estão na linha 1 e as 100 regiões metropolitanas estão nas linhas 2 a 101. As etapas a seguir podem ser usadas para selecionar uma amostra aleatória simples de 30 regiões metropolitanas.

Etapa 1. Digite =ALEATÓRIO( ) na célula C2
Etapa 2. Copie célula C2 para as células C3:C101
Etapa 3. Selecione qualquer célula da coluna C
Etapa 4. Dê um clique no botão **Classificar Crescente**

A amostra aleatória de 30 regiões metropolitanas aparecerá nas linhas 2 a 31 do conjunto de dados reorganizado. Os números aleatórios na Coluna C não são mais necessários e podem ser excluídos, se você quiser.

Estimação por Intervalo

ESTATÍSTICA NA PRÁTICA

FOOD LION*
Salisbury, Carolina do Norte

Fundada em 1957 com o nome de Food Town, a Food Lion é uma das maiores redes de supermercado dos Estados Unidos, com 1.200 lojas em 11 estados do sudeste e da região Mid-Atlantic.[1] A empresa vende mais de 24 mil diferentes produtos e oferece artigos de marca que têm publicidade em nível nacional e regional, bem como um crescente número de produtos com rótulo privado de alta qualidade manufaturados especialmente para a Food Lion. A empresa mantém sua liderança em preços baixos e garantia da qualidade pelas eficiências operacionais, como formatos de loja padronizados, projeto inovador de armazéns, instalações com uso eficiente da energia e sincronização de dados com os fornecedores. A Food Lion visa a um futuro de contínuas inovações, crescimento, liderança de preços e atendimento aos seus clientes.

Sendo integrante de um setor intensivo em inventários, a Food Lion decidiu adotar o método Ueps (último a entrar, primeiro a sair) de avaliação de inventários. Esse método compara os custos atuais com as receitas atuais, o que minimiza os efeitos das variações radicais de preço sobre os resultados de lucros e prejuízos. Além disso, o método Ueps reduz a receita líquida, diminuindo assim os impostos sobre a renda durante os períodos de inflação.

A Food Lion estabelece um índice Ueps para cada um dos sete agrupamentos de inventário: produtos de mercearia, papelaria e produtos domésticos, suprimentos para animais de estimação, saúde e beleza, laticínios, cigarros e tabaco e cervejas e vinhos. Por exemplo, o índice Ueps de 1,008 para o agrupamento pro-

* Os autores agradecem a Keith Cunningham, diretor do Departamento Fiscal da Food Lion, e a Bobby Harkey, da equipe de Contabilidade Fiscal da Food Lion por fornecer esta "Estatística na Prática".

[1] NT: *Mid-Atlantic. Adj.* – Caracterizado pela combinação de elementos, influências etc. – britânicos e norte-americanos. Diz-se dos estados norte-americanos que apresentam essas características.

dutos de mercearia indicaria que o valor de estoque dos produtos de mercearia da empresa aos custos atuais reflete um aumento de 0,8% em virtude da inflação no período mais recente de um ano.

O estabelecimento de um índice Ueps para cada agrupamento de inventário exige que a contagem de estoque de fim de ano referente a cada produto seja avaliada ao custo do fim de ano corrente e ao custo do fim de ano anterior. Para evitar o tempo e os gastos excessivos associados à contagem de estoques em todas as 1.200 lojas, a Food Lion seleciona uma amostra aleatória de 50 lojas.

São tomados os estoques físicos de fim de ano de cada uma das lojas da amostra. Os custos de cada item no ano corrente e do ano anterior são então utilizados para construir os índices Ueps necessários a cada agrupamento de inventário.

Em um ano recente, a estimativa amostral do índice Ueps referente ao agrupamento de inventário saúde e beleza foi de 1,015. Utilizando um grau de confiança de 95%, a Food Lion calculou a margem de erro de 0,006 para a estimativa amostral. Desse modo, o intervalo de 1,009 a 1,021 produziu uma estimação por intervalo do índice Ueps da população com um grau de confiança de 95%. Esse índice de precisão foi considerado muito bom.

Neste capítulo, você aprenderá a calcular a margem de erro associada a estimativas amostrais. Você também aprenderá a usar essa informação para construir e interpretar estimações por intervalo de uma média da população e de uma proporção da população.

No Capítulo 7, afirmamos que um estimador por ponto é uma estatística da amostra usada para estimar um parâmetro populacional. Por exemplo, a média $\bar{x}$ da amostra é um estimador por ponto da média populacional μ, e a proporção $\bar{p}$ da amostra é um estimador por ponto da proporção p da população. Uma vez que não se pode esperar que um estimador por ponto produza o valor exato do parâmetro populacional, uma **estimação por intervalo** freqüentemente é calculada adicionando-se e subtraindo-se um valor, denominado **margem de erro,** ao estimador por ponto. A forma geral de uma estimação por intervalo é a seguinte:

$$\text{Estimação por ponto} \pm \text{Margem de erro}$$

A finalidade de uma estimação por intervalo é fornecer informações sobre quão próximo o estimador por ponto, produzido pela amostra, está do valor do parâmetro populacional.

Neste capítulo, mostraremos como calcular estimativas por intervalo de uma média μ da população e de uma proporção p da população. A forma geral de uma estimação por intervalo de uma média populacional é:

$$\bar{x} \pm \text{Margem de erro}$$

Similarmente, a forma geral de uma estimação por intervalo de uma proporção populacional é:

$$\bar{p} \pm \text{Margem de erro}$$

As distribuições amostrais de $\bar{x}$ e $\bar{p}$ desempenham papéis fundamentais no cálculo dessas estimações por intervalo.

8.1 MÉDIA DA POPULAÇÃO: σ CONHECIDO

Para desenvolver uma estimação por intervalo da média de uma população, o desvio padrão σ da população ou o desvio padrão s da amostra deve ser usado para calcularmos a margem de erro. Na maioria das aplicações, σ não é conhecido, e usa-se s para calcular a margem de erro. Em algumas aplicações, entretanto, grandes quantidades de dados históricos relevantes estão disponíveis e podem ser utilizadas para calcular o desvio padrão da população antes de se fazer a amostragem. Igualmente, em aplicações de controle da qualidade nas quais se supõe que um processo esteja operando corretamente, ou "sob controle", é apropriado tratarmos o desvio padrão da população como conhecido. Referimo-nos a esse tipo de caso como aquele que apresenta **σ conhecido.** Nesta seção, apresentamos um exemplo em que é razoável tratarmos σ como conhecido e mostramos como construir uma estimação por intervalo para esse caso.

Semanalmente, a Lloyd's Department Store seleciona uma amostra aleatória simples de cem clientes para saber qual quantia eles gastam em cada ida às compras. Com x representando a quantia gasta em cada

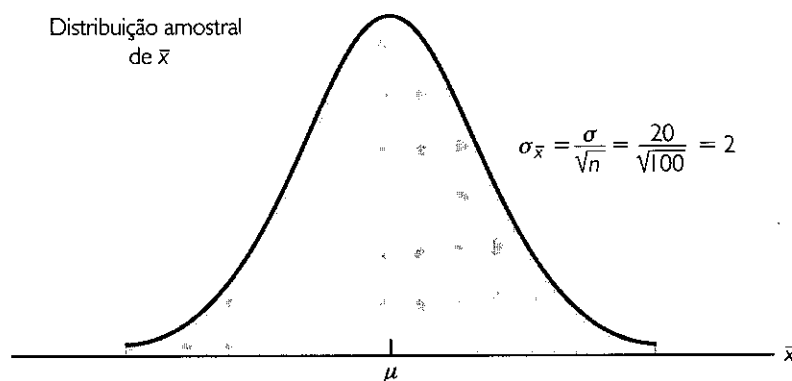
ida às compras, a média amostral $\bar{x}$ fornece uma estimação por ponto de μ, que é a quantia média gasta em cada ida às compras pela população de todos os clientes da empresa. A Lloyd's usa essa pesquisa semanal há vários anos. Baseando-se nos dados históricos, a empresa assume agora um valor conhecido de σ = US\$ 20 para o desvio padrão da população. Os dados históricos também indicam que a população segue uma distribuição normal.

ARQUIVO DA INTERNET Lloyd's

Durante a semana mais recente, a Lloyd's pesquisou 100 clientes (n = 100) e obteve a média $\bar{x}$ da amostra = US\$ 82,00. A quantia média gasta pela amostra fornece uma estimação por ponto da quantia média gasta pela população em cada ida às compras. Na discussão a seguir, mostramos como calcular a margem de erro dessa estimação e como desenvolver uma estimação por intervalo da média da população.

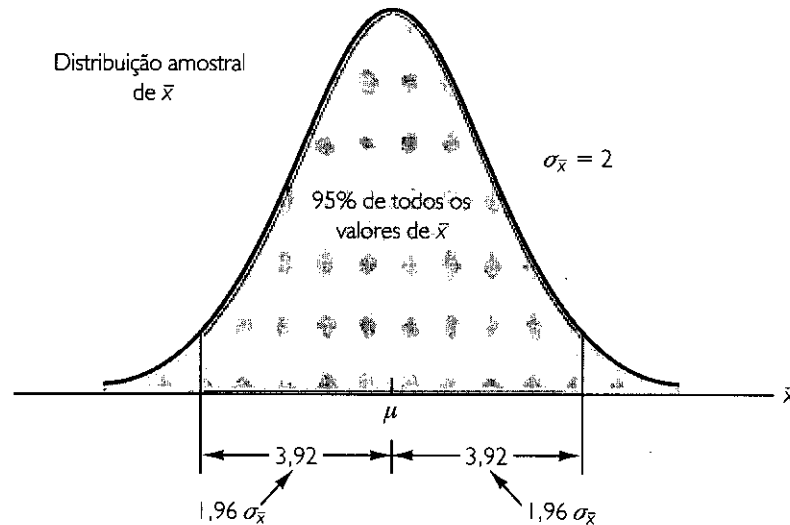
Margem de Erro e a Estimação por Intervalo

No Capítulo 7, mostramos que a distribuição amostral de $\bar{x}$ pode ser usada para calcularmos a probabilidade de $\bar{x}$ estar dentro de determinada distância de μ. No exemplo da Lloyd's, os dados históricos mostram que a população das quantias gastas está normalmente distribuída, com um desvio padrão σ = 20. Então, utilizando o que aprendemos no Capítulo 7, podemos concluir que a distribuição amostral de $\bar{x}$ segue uma distribuição normal, com um erro padrão de $\sigma_{\bar{x}} = \sigma/\sqrt{n} = 20/\sqrt{100} = 2$. Essa distribuição amostral é mostrada na Figura 8.1.[2]

Figura 8.1 Distribuição amostral da quantia média que os integrantes da amostra gastaram, obtida de amostras aleatórias simples de 100 clientes

Uma vez que a distribuição amostral mostra como os valores de $\bar{x}$ estão distribuídos nas proximidades da média populacional μ, a distribuição amostral de $\bar{x}$ fornece informações sobre as possíveis diferenças entre $\bar{x}$ e μ.

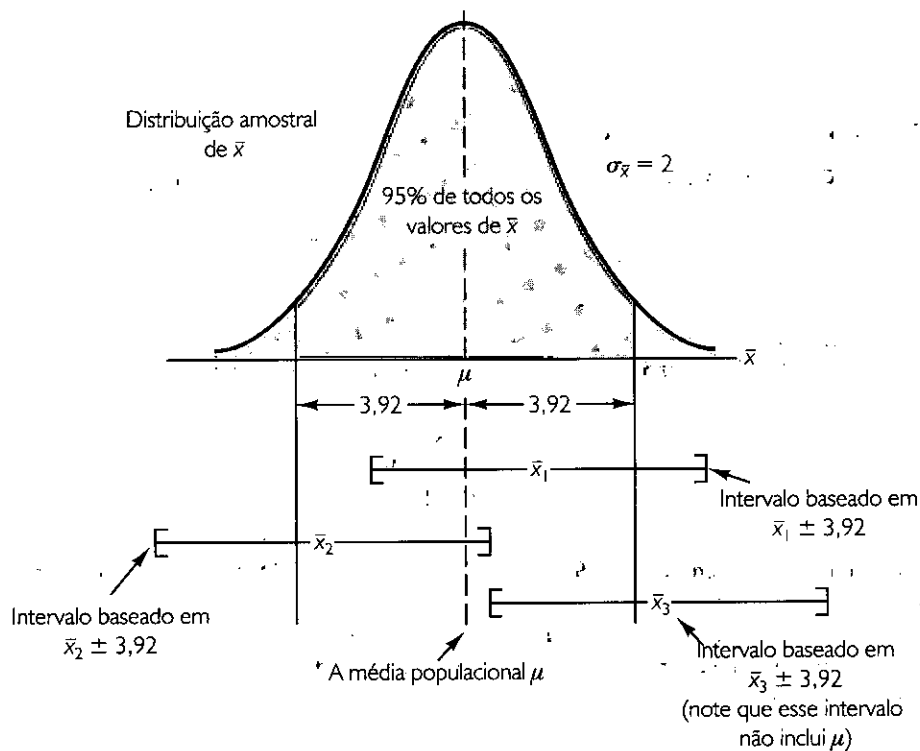
Usando a tabela de áreas da distribuição normal padrão, descobrimos que 95% dos valores de qualquer variável aleatória normalmente distribuída estão dentro de ±1,96 desvio padrão da média. Desse modo, quando a distribuição amostral de $\bar{x}$ está normalmente distribuída, 95% dos valores de $\bar{x}$ devem estar dentro de $\pm 1{,}96\sigma_{\bar{x}}$ da média μ. No exemplo da Lloyd's, sabemos que a distribuição amostral de $\bar{x}$ está normalmente distribuída, com um erro padrão de $\sigma_{\bar{x}}=2$. Uma vez que $\pm 1{,}96\ \sigma_{\bar{x}} = 1{,}96(2) = 3{,}92$, podemos concluir que 95% de todos os valores de $\bar{x}$, obtidos usando-se um tamanho de amostra n = 100, estarão dentro de ±3,92 da média populacional μ. Veja a Figura 8.2.

[2] Usamos o fato de que a população de quantias gastas tem uma distribuição normal para concluir que a distribuição amostral de $\bar{x}$ tem uma distribuição normal. Se a população não tivesse uma distribuição normal, poderíamos recorrer ao teorema do limite central e ao tamanho da amostra n = 100 para concluir que a distribuição amostral de $\bar{x}$ é aproximadamente normal. Em qualquer um dos casos, a distribuição amostral de $\bar{x}$ se assemelharia à que é apresentada na Figura 8.1.

Figura 8.2 Distribuição amostral de $\bar{x}$ indicando a localização das médias amostrais que estão dentro de 3,92 de μ

Na introdução deste capítulo, dissemos que a forma geral da estimação por intervalo da média populacional μ é ± a margem de erro. No exemplo da Lloyd's, suponha que definamos a margem de erro em 3,92 para calcular a estimação por intervalo de μ usando $\bar{x} \pm 3{,}92$. Para produzirmos uma interpretação para essa estimação por intervalo, vamos considerar os valores de $\bar{x}$ que poderiam ser obtidos se tivéssemos tomado três *diferentes* amostras aleatórias simples, cada uma das quais consistindo em 100 clientes da Lloyd's. A primeira média amostral poderia assumir o valor apresentado como $\bar{x}_1$ na Figura 8.3. Nesse caso, a Figura 8.3 mostra que o intervalo formado ao subtrair-se 3,92 de $\bar{x}_1$ e adicionar-se 3,92 a $\bar{x}_1$ inclui a média populacional μ. Considere agora o que acontece se a segunda média amostral assumir o valor apresentado como $\bar{x}_2$ na Figura 8.3. Não obstante essa média amostral diferir da primeira média amostral, notamos que o intervalo formado ao subtrair-se 3,92 de $\bar{x}_2$ e adicionar-se 3,92 a $\bar{x}_2$ também inclui a média populacional μ. Entretanto, considere o que acontece se a terceira média amostral assumir o valor apresentado como $\bar{x}_3$ na Figura 8.3. Nesse caso, o intervalo formado ao subtrair-se 3,92 de $\bar{x}_3$ e adicionar-se 3,92 a $\bar{x}_3$ não inclui a média populacional μ. Uma vez que $\bar{x}_3$ se situa na cauda superior (*upper tail*) da distribuição amostral e tem um afastamento maior que 3,92 de μ, subtrair ou adicionar 3,92 a $\bar{x}_3$ forma um intervalo que não inclui μ.

Qualquer média amostral $\bar{x}$ que esteja dentro da área com sombreamento mais escuro da Figura 8.3 fornecerá um intervalo que contém a média populacional μ. Visto que 95% de todas as médias amostrais possíveis estão na área com sombreamento mais escuro, 95% de todos os intervalos formados subtraindo-se 3,92 de $\bar{x}$ ou adicionando-se 3,92 a $\bar{x}$ incluirão a média populacional μ.

Lembre-se de que durante a semana mais recente, a equipe de garantia da qualidade da Lloyd's pesquisou 100 clientes e obteve uma média amostral de quantias gastas de $\bar{x} = 82$. Usando $\bar{x} \pm 3{,}92$ para construir a estimação por intervalo, obtemos 82 ± 3,92. Assim, a estimação por intervalo de μ específica baseada nos dados da semana mais recente é igual a 82 – 3,92 = 78,08 a 82 + 3,92 = 85,92. Uma vez que 95% de todos os intervalos construídos usando-se $\bar{x} \pm 3{,}92$ conterão a média populacional, dizemos que temos 95% de confiança em que o intervalo 78,08 a 85,92 inclui a média populacional μ. Dizemos que esse intervalo foi estabelecido com o **grau de confiança** de 95%. O valor 0,95 denomina-se **coeficiente de confiança**, e o intervalo de 78,08 a 85,92 é chamado **intervalo de confiança** de 95%.

Figura 8.3 Intervalos formados a partir das médias amostrais selecionadas nas posições x_1, x_2 e x_3

Com a margem de erro dada por $z_{\alpha/2}(\sigma/\sqrt{n})$, a forma geral de uma estimação por intervalo de uma média populacional para o caso de σ conhecido é a seguinte:

ESTIMAÇÃO POR INTERVALO DE UMA MÉDIA POPULACIONAL: σ CONHECIDO

$$\bar{x} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \tag{8.1}$$

em que $(1 - \alpha)$ é o coeficiente de confiança e $z_{\alpha/2}$ é o valor z que produz uma área de $\alpha/2$ na cauda superior da distribuição normal padrão de probabilidade.

Essa discussão nos dá subsídios para compreender por que o intervalo é chamado intervalo de confiança de 95%.

Vamos usar a Equação 8.1 para construir um intervalo de confiança de 95% para o exemplo da Lloyd's. Para um intervalo de confiança de 95%, o coeficiente de confiança é $(1 - \alpha) = 0,95$ e, desse modo, $\alpha = 0,05$. Usando as tabelas de áreas da distribuição normal padrão, uma área de $\alpha/2 = 0,05/2 = 0,025$ na cauda superior produz $z_{0,025} = 1,96$. Com a média amostral de $\bar{x} = 82$, $\sigma = 20$ e um tamanho de amostra $n = 100$ para o caso da Lloyd's, obtemos

$$82 \pm 1,96\frac{20}{\sqrt{100}}$$

$$82 \pm 3,92$$

Dessa forma, usando a Equação 8.1, a margem de erro é 3,92, e o intervalo de confiança de 95% é $82 - 3,92 = 78,08$ a $82 + 3,92 = 85,92$.

Não obstante um grau de confiança de 95% freqüentemente ser usado, outros graus de confiança, por exemplo, 90% e 99%, podem ser considerados. A Tabela 8.1 apresenta os valores de $z_{\alpha/2}$ correspondentes aos graus de confiança mais comumente utilizados. Usando esses valores e a Equação 8.1, o intervalo de confiança de 90% para o exemplo da Lloyd's é:

$$82 \pm 1{,}645 \frac{20}{\sqrt{100}}$$
$$82 \pm 3{,}29$$

Tabela 8.1 Valores de $z_{\alpha/2}$ correspondentes aos graus de confiança mais comumente utilizados

Grau de Confiança	α	$\alpha/2$	$z_{\alpha/2}$
90%	0,10	0,05	1,645
95%	0,05	0,025	1,960
99%	0,01	0,005	2,576

Assim, com um grau de confiança de 90%, a margem de erro é 3,29 e o intervalo de confiança é 82 – 3,29 =78,71 a 82 + 3,29 =85,29. Similarmente, o intervalo de confiança de 99% é:

$$82 \pm 2{,}576 \frac{20}{\sqrt{100}}$$
$$82 \pm 5{,}15$$

Portanto, com um grau de confiança de 99%, a margem de erro é 5,15 e o intervalo de confiança é 82 – 5,15 = 76,85 a 82 + 5,15 = 87,15.

Combinando os resultados correspondentes aos graus de confiança de 90%, 95% e 99%, observamos que, para termos um grau de confiança mais elevado, a margem de erro e, portanto, a amplitude do intervalo de confiança devem ser maiores.

Conselho Prático

Se a população segue uma distribuição normal, o intervalo de confiança produzido pela Equação 8.1 é exato. Em outras palavras, se a Equação 8.1 fosse usada repetidamente para gerar intervalos de confiança de 95%, exatamente 95% dos intervalos gerados conteriam a média da população. Se a população não segue uma distribuição normal, o intervalo de confiança produzido pela Equação 8.1 será aproximado. Nesse caso, a qualidade da aproximação depende tanto da distribuição da população como do tamanho da amostra.

Na maioria das aplicações, um tamanho de amostra $n \geq 30$ é adequado quando se usa a Equação 8.1 para desenvolver uma estimação por intervalo de uma média populacional. Se a população não está normalmente distribuída, mas é aproximadamente simétrica, pode-se esperar que tamanhos de amostra pequenos, até mesmo de 15, produzam bons intervalos de confiança aproximados. Com tamanhos de amostra menores, a Equação 8.1 somente deve ser usada se o analista acreditar, ou estiver disposto a supor, que a distribuição populacional seja, no mínimo, aproximadamente normal.

NOTAS E COMENTÁRIOS

1. O procedimento de estimação por intervalo discutido nesta seção baseia-se no pressuposto de que o desvio padrão σ seja conhecido. Por "σ conhecido" queremos dizer que há dados históricos ou outras informações disponíveis que nos permitem obter uma boa estimativa do desvio padrão da população antes de tomarmos a mostra que será usada para desenvolver uma estimativa da média populacional. Então, tecnicamente, não queremos dizer que σ seja, de fato, conhecido com certeza. Simplesmente, queremos dizer que obtivemos uma boa estimativa do desvio padrão antes de fazer a amostragem e, desse modo, não usaremos a mesma amostra para estimar tanto a média populacional como o desvio padrão da população.
2. O tamanho n da amostra aparece no denominador da expressão de estimação por intervalo (Equação 8.1). Assim, se uma amostra em particular produzir um intervalo demasiadamente amplo para ter uso prático, talvez queiramos considerar aumentar o tamanho da amostra. Com n no denominador, um tamanho de amostra maior produzirá uma margem de erro menor, um intervalo mais estreito e uma precisão maior. O procedimento para determinar o tamanho de uma amostra aleatória simples necessária para se obter a precisão desejada será discutido na Seção 8.3.

Exercícios

Métodos

1. Uma amostra aleatória simples de 40 itens resultou em uma média amostral 25. O desvio padrão da população é $\sigma = 5$.
   a. Qual é o erro padrão da média, $\sigma_{\bar{x}}$?
   b. Para um grau de confiança de 95%, qual é a margem de erro?
2. Uma amostra aleatória simples de 50 itens de uma população, com $\sigma = 6$, resultou em uma média amostral igual a 32.
   a. Forneça um intervalo de confiança de 90% para a média populacional.
   b. Estime um intervalo de confiança de 95% para a média populacional.
   c. Providencie um intervalo de confiança de 99% para a média populacional.

AUTOTESTE

3. Uma amostra aleatória simples de 60 itens resultou em uma média amostral igual a 80. O desvio padrão σ da população é igual a 15.
   a. Calcule o intervalo de confiança de 95% para a média populacional.
   b. Suponha que a mesma média amostral tenha sido obtida de uma amostra de 120 itens. Forneça um intervalo de confiança de 95% da média populacional.
   c. Qual é o efeito de um tamanho de amostra maior sobre a estimação por intervalo?
4. Sabe-se que o intervalo de confiança de 95% de uma média populacional é de 152 a 160. Se $\sigma = 15$, qual tamanho de amostra foi utilizado nesse estudo?

Aplicações

5. Em um esforço para estimar a quantia média que cada cliente gasta por jantar em um grande restaurante de Atlanta, foram coletados dados de uma amostra de 49 clientes. Suponha um desvio padrão de US$ 5,00 para a população.
   a. Para um grau de confiança de 95%, qual é a margem de erro?
   b. Se a média amostral é US$ 24,80, qual é o intervalo de confiança de 95% para a média populacional?
6. A Nielsen Media Research relatou que o tempo médio que as famílias passam assistindo à televisão, no período das 8h às 11h da noite, é de 8,5 horas por semana (*The World Almanac 2003*). Dado um tamanho de amostra de 300 famílias e um desvio padrão σ da população igual a 3,5 horas, qual é a estimação por intervalo de confiança de 95% da média de tempo que as pessoas assistem à televisão durante o período das 8h às 11h da noite?

AUTOTESTE

7. Uma pesquisa de pequenos negócios com *websites* revelou que a quantia média gasta em um site era de US$ 11.500 por ano (*Fortune*, 5 de março de 2001). Dada uma amostra de 60 negócios e um desvio padrão σ da população igual a US$ 4 mil, qual é a margem de erro? Use 95% de confiança. O que você recomendaria se o estudo demandasse uma margem de erro de US$ 500?
8. O National Quantity Research Center da Universidade de Michigan publica uma medida trimestral das opiniões dos consumidores sobre produtos e serviços (*The Wall Street Journal*, 18 de fevereiro de 2003). Uma pesquisa de dez restaurantes do grupo Fast Food/Pizza revelou que a média amostral de satisfação do cliente tinha um índice igual a 71. Dados históricos indicam que o desvio padrão populacional do índice era relativamente estável, com $\sigma = 5$.
   a. Qual suposição o pesquisador estaria disposto a fazer se fosse desejada uma margem de erro?
   b. Usando um grau de confiança de 95%, qual é a margem de erro?
   c. Qual é a margem de erro se for desejado um grau de confiança igual a 99%?
9. O *undergraduate grade point average* (GPA)[3] para estudantes matriculados nas melhores escolas de pós-graduação em Administração foi de 3,37 (*Best Graduate Schools, U.S. News and World Report*, 2001). Suponha que essa estimativa tenha se baseado em uma amostra de 120 estudantes matriculados nas melhores escolas. Usando-se os dados de anos anteriores, o desvio padrão da população pode

[3] NT: GPA: *Educ.* – Média de notas, média escolar. Uma medida numérica do rendimento acadêmico baseada no cálculo do número de créditos e notas obtidas em todas as matérias até o presente. Baseia-se em uma escala de 0 a 4 (Estados Unidos).

ser considerado conhecido, com σ =0,28. Qual é a estimação por intervalo de confiança de 95% da GPA para estudantes matriculados nas principais escolas de pós-graduação em Administração?

10. A revista *Playbill* divulgou que a renda familiar anual média de seus leitores é igual a US$ 119.155 (*Playbill*, dezembro de 2003). Suponha que essa estimativa da renda familiar anual média se baseie em uma amostra de 80 famílias; com base em estudos passados, sabe-se que o desvio padrão da população é σ =US$ 30.000.
    a. Desenvolva uma estimação por intervalo de confiança de 90% para a média populacional.
    b. Estabeleça uma estimação por intervalo de confiança de 95% para a média populacional.
    c. Determine uma estimação por intervalo de confiança de 99% para a média populacional.
    d. Discuta o que acontece à amplitude do intervalo de confiança quando o grau de confiança é aumentado. Esse resultado parece razoável? Explique.

8.2 MÉDIA DA POPULAÇÃO: σ DESCONHECIDO

Quando desenvolvemos a estimação por intervalo de uma média populacional, geralmente não temos uma boa estimativa do desvio padrão da população. Nesses casos, precisamos usar a mesma amostra para estimar μ e σ. Essa situação representa o caso que apresenta σ **desconhecido**. Quando s é usado para estimar σ, a margem de erro e a estimação por intervalo da média populacional baseiam-se em uma distribuição de probabilidade conhecida como **distribuição *t***. Não obstante o desenvolvimento matemático da distribuição *t* basear-se na suposição de uma distribuição normal para a população da qual extraímos a amostra, as pesquisas mostram que a distribuição *t* pode ser aplicada de maneira bem-sucedida em muitas situações em que a população se desvia significativamente da normal. Posteriormente, nesta seção, apresentaremos diretrizes para se usar a distribuição *t* se a população não estiver normalmente distribuída.

William Sealy Gosset, escritor que usava o pseudônimo "Student", é o descobridor da distribuição t. Gosset, graduado em Matemática pela Universidade de Oxford, trabalhava para a Guinness Brewery (Cervejarias Guinness), em Dublin, Irlanda. Ele desenvolveu a distribuição t enquanto trabalhava em materiais de pequena escala e experimentos com temperatura.

A distribuição *t* é uma família de distribuições de probabilidade similares, com uma distribuição *t* específica que depende de um parâmetro conhecido como **grau de liberdade**. A distribuição *t* com um grau de liberdade é única, como o é a distribuição *t* com dois graus de liberdade, com três graus de liberdade e assim por diante. À medida que o número de graus de liberdade aumenta, a diferença entre a distribuição *t* e a distribuição normal padrão torna-se cada vez menor. A Figura 8.4 apresenta distribuições *t* com valores de 10 e 20 graus de liberdade e suas relações com a distribuição normal de probabilidade. Note que uma distribuição *t* com mais graus de liberdade exibe menos variabilidade e se assemelha mais estreitamente à distribuição normal padrão. Note também que a média da distribuição *t* é zero.

Figura 8.4 Comparação da distribuição normal padrão com distribuições t que têm 10 e 20 graus de liberdade

Colocamos um subscrito em t para indicar a área na cauda superior (*upper tail*) da distribuição t. Por exemplo, do mesmo modo que usamos $z_{0,025}$ para indicar o valor z que produz uma área de 0,025 na cauda superior de uma distribuição normal padrão, usaremos $t_{0,025}$ para indicar o valor t que produz uma área de 0,025 na cauda superior de uma distribuição t. Em geral, usaremos a notação $t_{\alpha/2}$ para representar um valor t com uma área de $\alpha/2$ na cauda superior da distribuição t. Veja a Figura 8.5.

À medida que os graus de liberdade aumentam, a distribuição t se aproxima da distribuição normal padrão.

A Tabela 8.2 é uma tabela da distribuição t. Cada linha da tabela corresponde a uma distribuição t distinta, com os graus de liberdade correspondentes. Por exemplo, para uma distribuição t com 10 graus de liberdade, $t_{0,025} = 2,228$. Similarmente, para uma distribuição t com 20 graus de liberdade, $t_{0,025} = 2,086$. À medida que os graus de liberdade continuam a crescer, $t_{0,025}$ se aproxima de $z_{0,025} = 1,96$. De fato, os valores z da distribuição normal padrão podem ser encontrados na linha de graus de liberdade infinitos (rotulada com ∞) da tabela de distribuições t. Se o grau de liberdade ultrapassar 100, a linha de graus de liberdade infinitos pode ser usada para aproximar o valor real t; em outras palavras, para mais de 100 graus de liberdade, o valor z normal padrão fornece uma boa aproximação ao valor t. A Tabela 2 do Apêndice B é uma tabela de distribuições t mais extensa, com todos os graus de liberdade, de 1 a 100, inclusive.

Margem de Erro e a Estimação por Intervalo

Na Seção 8.1, mostramos que a estimação por intervalo de uma média populacional para o caso de σ conhecido é:

$$\bar{x} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

Para calcular uma estimação por intervalo de μ para o caso de σ desconhecido, o desvio padrão σ da amostra é usado para estimar σ, e $z_{\alpha/2}$ é substituído pelo valor da distribuição t, $t_{\alpha/2}$. A margem de erro é dada então por $t_{\alpha/2}s/\sqrt{n}$.

Figura 8.5 Distribuição t com a área, ou probabilidade, $\alpha/2$ na cauda superior

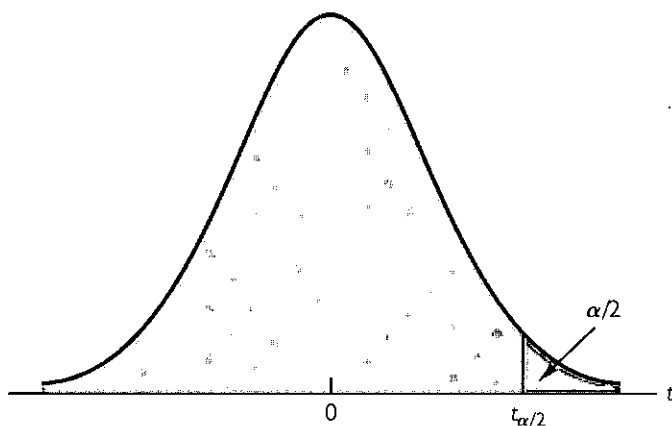
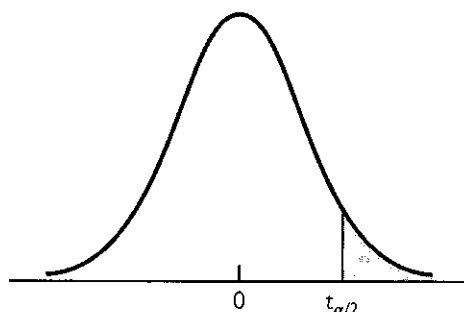



Tabela 8.2 Tabela de distribuição t de uma área $\alpha/2$ na cauda superior. Exemplo: com 10 graus de liberdade, o valor t que produz uma área de 0,025 na cauda superior é $t_{0,025} = 2,228$

Graus de Liberdade	Área na Cauda Superior					
	0,20	**0,10**	**0,05**	**0,025**	**0,01**	**0,005**
1	1,376	3,078	6,314	12,706	31,821	63,656
2	1,061	1,886	2,920	4,303	6,965	9,925
3	0,978	1,638	2,353	3,182	4,541	5,841
4	0,941	1,533	2,132	2,776	3,747	4,604
5	0,920	1,476	2,015	2,571	3,365	4,032
6	0,906	1,440	1,943	2,447	3,143	3,707
7	0,896	1,415	1,895	2,365	2,998	3,499
8	0,889	1,397	1,860	2,306	2,896	3,355
9	0,883	1,383	1,833	2,262	2,821	3,250
10	0,879	1,372	1,812	2,228	2,764	3,169
11	0,876	1,363	1,796	2,201	2,718	3,106
12	0,873	1,356	1,782	2,179	2,681	3,055
13	0,870	1,350	1,771	2,160	2,650	3,012
14	0,868	1,345	1,761	2,145	2,624	2,977
15	0,866	1,341	1,753	2,131	2,602	2,947
16	0,865	1,337	1,746	2,120	2,583	2,921
17	0,863	1,333	1,740	2,110	2,567	2,898
18	0,862	1,330	1,734	2,101	2,552	2,878
19	0,861	1,328	1,729	2,093	2,539	2,861
20	0,860	1,325	1,725	2,086	2,528	2,845
21	0,859	1,323	1,721	2,080	2,518	2,831
22	0,858	1,321	1,717	2,074	2,508	2,819
23	0,858	1,319	1,714	2,069	2,500	2,807
24	0,857	1,318	1,711	2,064	2,492	2,797
25	0,856	1,316	1,708	2,060	2,485	2,787
26	0,856	1,315	1,706	2,056	2,479	2,779
27	0,855	1,314	1,703	2,052	2,473	2,771
28	0,855	1,313	1,701	2,048	2,467	2,763
29	0,854	1,311	1,699	2,045	2,462	2,756
30	0,854	1,310	1,697	2,042	2,457	2,750
40	0,851	1,303	1,684	2,021	2,423	2,704
50	0,849	1,299	1,676	2,009	2,403	2,678
60	0,848	1,296	1,671	2,000	2,390	2,660
80	0,846	1,292	1,664	1,990	2,374	2,639
100	0,845	1,290	1,660	1,984	2,364	2,626
∞	0,842	1,282	1,645	1,960	2,326	2,576

Nota: Uma tabela mais extensa é apresentada na Tabela 2 do Apêndice B.

Com essa margem de erro, a expressão geral de uma estimação por intervalo de uma média populacional quando σ é desconhecido é a seguinte:

ESTIMAÇÃO POR INTERVALO DE UMA MÉDIA POPULACIONAL: σ DESCONHECIDO

$$\bar{x} \pm t_{\alpha/2} \frac{s}{\sqrt{n}} \tag{8.2}$$

em que s é o desvio padrão da amostra, $(1 - \alpha)$ é o coeficiente de confiança e $t_{\alpha/2}$ é o valor t que produz uma área igual a $\alpha/2$ na cauda superior da distribuição t, com $n - 1$ graus de liberdade.

A razão pela qual o número de graus de liberdade associado ao valor t na Equação 8.2 é $n - 1$ refere-se ao uso de s como uma estimativa do desvio padrão s da população. A expressão do desvio padrão da amostra é:

$$s = \sqrt{\frac{\Sigma(x_i - \bar{x})^2}{n - 1}}$$

Os graus de liberdade referem-se ao número de informações independentes que entram no cálculo de $\Sigma(x_i - \bar{x})^2$. As n informações independentes envolvidas no cálculo de $\Sigma(x_i - \bar{x})^2$ são as seguintes: $x_1 - \bar{x}$, $x_2 - \bar{x}, \ldots, x_n - \bar{x}$. Na Seção 3.2, indicamos que $\Sigma(x_i - \bar{x}) = 0$ para qualquer conjunto de dados. Desse modo, somente $n - 1$ dos $x_i - \bar{x}$ valores são independentes; ou seja, se conhecermos $n - 1$ dos valores, o valor restante pode ser determinado de maneira exata, usando-se a condição de que a soma dos $x_i - \bar{x}$ valores deve ser igual a 0. Assim, $n - 1$ é o número de graus de liberdade associados a $\Sigma(x_i - \bar{x})^2$ e, portanto, o número de graus de liberdade da distribuição t na Equação 8.2.

Para ilustrar o procedimento de estimação por intervalo para o caso de s desconhecido, consideraremos um estudo idealizado para estimar a média dos débitos de cartão de crédito da população de famílias norte-americanas. Uma amostra de n =85 famílias forneceu os saldos de cartões de crédito mostrados na Tabela 8.3. Para essa situação, nenhuma estimativa anterior do desvio padrão s da população está disponível.

Tabela 8.3 Saldos de cartões de crédito de uma amostra de 85 famílias

9.619	5.994	3.344	7.888	7.581	9.980
5.364	4.652	13.627	3.091	12.545	8.718
8.348	5.376	968	943	7.959	8.452
7.348	5.998	4.714	8.762	2.563	4.935
381	7.530	4.334	1.407	6.787	5.938
2.998	3.678	4.911	6.644	5.071	5.266
1.686	3.581	1.920	7.644	9.536	10.658
1.962	5.625	3.780	11.169	4.459	3.910
4.920	5.619	3.478	7.979	8.047	7.503
5.047	9.032	6.185	3.258	8.083	1.582
6.921	13.236	1.141	8.660	2.153	
5.759	4.447	7.577	7.511	8.003	
8.047	609	4.667	14.442	6.795	
3.924	414	5.219	4.447	5.915	
3.470	7.636	6.416	6.550	7.164	

ARQUIVO DA INTERNET
Balance

Sendo assim, dados amostrais precisam ser utilizados para se estimar tanto a média populacional como o desvio padrão da população. Usando-se os dados da Tabela 8.3, calculamos a média amostral $\bar{x}$ = US$ 5.900 e o desvio padrão s da amostra = US$ 3.058. Com 95% de confiança e $n - 1 = 84$ graus de liberdade, a Tabela 2 do Apêndice B fornece $t_{0,025} = 1,989$. Agora, podemos usar a Equação 8.2 para calcular uma estimação por intervalo da média populacional:

$$5.900 \pm 1,989 \frac{3.058}{\sqrt{85}}$$

$$5.900 \pm 660$$

A estimação por ponto da média populacional é US$ 5.900, a margem de erro é US$ 660 e o intervalo de confiança de 95% é de 5.900 – 660 =US$ 5.240 a 5.900 + 660 = US$ 6.560. Desse modo, temos 95% de confiança em que a média dos saldos de cartão de crédito da população de todas as famílias está entre US$ 5.240 e US$ 6.560.

Os procedimentos usados pelo Minitab e pelo Excel para desenvolver intervalos de confiança para uma média populacional são descritos nos Apêndices 8.1 e 8.2. Em relação ao estudo de saldos de cartão de crédito das famílias norte-americanas, os resultados do procedimento de estimação por intervalo do Minitab são mostrados na Figura 8.6. A amostra de 85 famílias produz uma média amostral de estratos de cartão de crédito igual a US$ 5.900, desvio padrão de US$ 3.058 e (após o arredondamento) uma estimativa do erro padrão da média igual a US$ 332, e um intervalo de confiança de 95% igual a US$ 5.240 até US$ 6.560.

Conselho Prático

Se a população segue uma distribuição normal, o intervalo de confiança produzido pela Equação 8.2 é exato e pode ser usado para qualquer tamanho de amostra. Se a população não segue uma distribuição normal, o intervalo de confiança produzido pela Equação 8.2 será aproximado. Nesse caso, a qualidade da aproximação depende tanto da distribuição da população como do tamanho da amostra.

Tamanhos de amostra maiores são necessários se a distribuição da população for altamente inclinada ou se incluir pontos fora da curva.

Na maioria das aplicações, um tamanho de amostra $n \geq 30$ é adequado quando se usa a Equação 8.2 para desenvolver uma estimação por intervalo de uma média populacional. Entretanto, se a distribuição populacional for altamente inclinada ou se contiver pontos fora da curva, a maioria dos estatísticos recomendaria aumentar o tamanho da amostra para 50 ou mais. Se a população não está normalmente distribuída, mas é mais ou menos simétrica, pode-se esperar que tamanhos de amostra tão pequenos quanto 15 produzam bons intervalos de confiança aproximados. Com tamanhos de amostra menores, a Equação 8.2 somente deve ser usada se o analista acreditar, ou estiver disposto a supor, que a distribuição populacional seja, no mínimo, aproximadamente normal.

Como Usar uma Amostra Pequena

No exemplo a seguir, desenvolvemos uma estimação por intervalo de uma média populacional quando o tamanho da amostra é pequeno. Conforme já observamos, um entendimento da distribuição populacional torna-se um fator importante ao decidir se o procedimento de estimação por intervalo produz resultados aceitáveis.

A Scheer Industries está considerando usar um novo programa auxiliado por computador para treinar os empregados do setor de manutenção a fazer reparos nas máquinas. A fim de avaliar plenamente o programa, o diretor do departamento de manufatura solicitou uma estimativa do tempo médio populacional necessário para que os empregados do setor de manutenção concluam o treinamento auxiliado por computador.

ARQUIVO DA INTERNET
Balance

Figura 8.6 Intervalo de confiança do Minitab para a pesquisa de saldos de cartão de crédito

Variable	N	Mean	StDev	SE Mean	95% CI
Balance	85	5.900,00	3.058,00	331,69	(5.240,40, 6.559,60)

Tabela 8.4 Tempo de treinamento, em dias, correspodente à amostra de 20 empregados da Scheer Industries

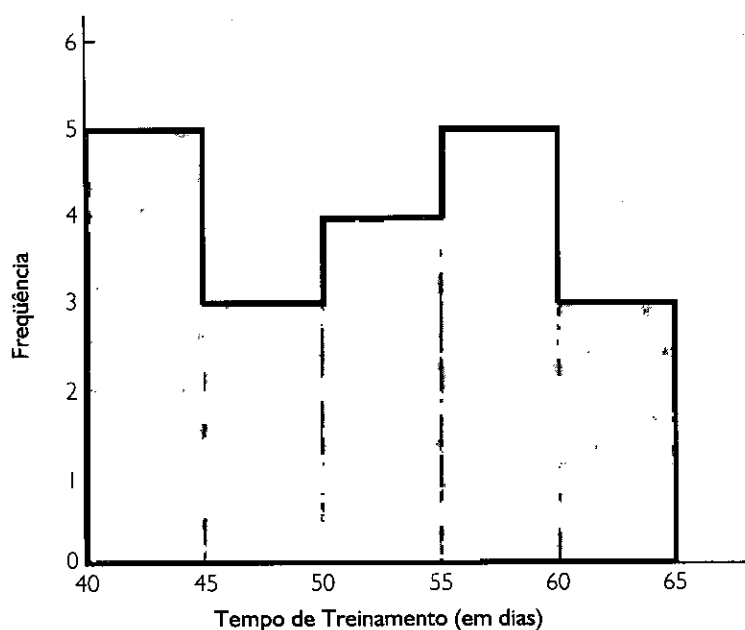
52	59	54	42
44	50	42	48
55	54	60	55
44	62	62	57
45	46	43	56

Uma amostra de 20 empregados é selecionada, tendo cada empregado da amostra concluído o programa de treinamento. Os dados sobre o tempo de treinamento, em dias, correspondentes aos 20 empregados, são mostrados na Tabela 8.4. Um histograma dos dados da amostra é apresentado na Figura 8.7. O que se pode dizer a respeito da distribuição da população com base nesse histograma? Primeiro, os dados da amostra não sustentam a conclusão de que a distribuição da população seja normal, ainda que não vejamos nenhuma evidência de inflexão ou de pontos fora da curva. Portanto, usando as diretrizes apresentadas na subseção anterior, concluímos que uma estimação por intervalo baseada na distribuição t parece aceitável para a amostra de 20 empregados.

Continuamos a calcular a média amostral e o desvio padrão da amostra da seguinte maneira:

$$\bar{x} = \frac{\Sigma x_i}{n} = \frac{1030}{20} = 51{,}5 \text{ dias}$$

$$s = \sqrt{\frac{\Sigma(x_i - \bar{x})^2}{n-1}} = \sqrt{\frac{889}{20-1}} = 6{,}84 \text{ dias}$$

Figura 8.7 Histograma dos tempos de treinamento da amostra da Scheer Industries

Para um intervalo de confiança de 95%, usamos a Tabela 8.2 e $n - 1 = 19$ graus de liberdade para obter $t_{0,025} = 2{,}093$. A Equação 8.2 fornece a estimação por intervalo da média populacional.

$$51{,}5 \pm 2{,}093\left(\frac{6{,}84}{\sqrt{20}}\right)$$

$$51{,}5 \pm 3{,}2$$

A estimação por ponto da média populacional é igual a 51,5 dias. A margem de erro é 3,2 dias e o intervalo de confiança de 95% é de 51,5 – 3,2 = 48,3 dias a 51,5 + 3,2 = 54,7 dias.

Usar um histograma dos dados da amostra para conhecer a distribuição de uma população nem sempre é conclusivo, mas, em muitos casos, constitui a única informação disponível. O histograma, juntamente com o julgamento da parte do analista, freqüentemente pode ser utilizado para decidir se a Equação 8.2 pode ser usada para desenvolver a estimação por intervalo.

Resumo dos Procedimentos de Estimação

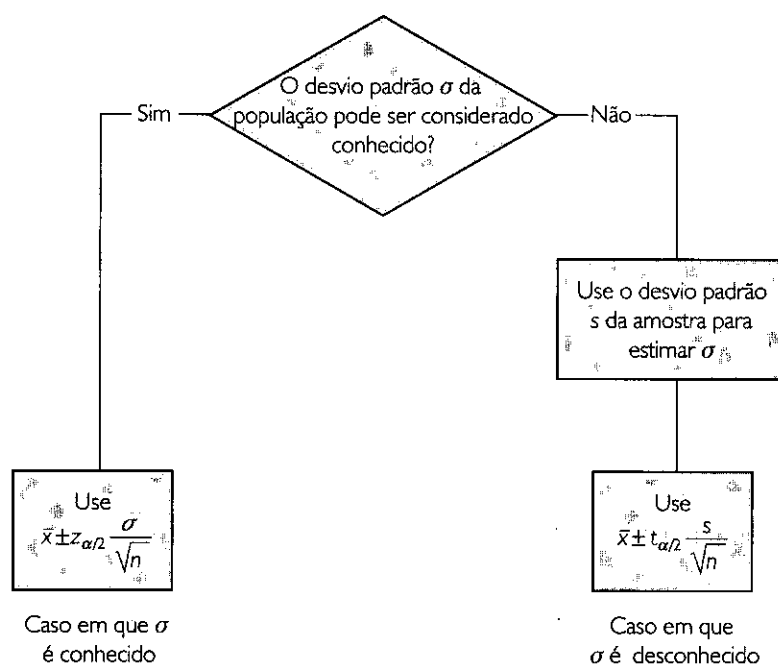
Apresentamos duas abordagens ao desenvolvimento de uma estimação por intervalo de uma média populacional. Para o caso de σ conhecido, o σ e a distribuição normal padrão são utilizados na Equação 8.1 para calcular a margem de erro e para desenvolver a estimação por intervalo. Para o caso de σ desconhecido, o desvio padrão s da amostra e a distribuição t são utilizados na Equação 8.2 para calcular a margem de erro e para desenvolver a estimação por intervalo.

Um resumo dos procedimentos de estimação por intervalo referente aos dois casos é apresentado na Figura 8.8. Na maioria das aplicações, um tamanho de amostra $n \geq 30$ é adequado. Entretanto, se a população tiver uma distribuição normal ou aproximadamente normal, tamanhos de amostra menores poderão ser usados. Para o caso de σ desconhecido, um tamanho de amostra $n \geq 50$ é recomendado quando se acredita que a distribuição populacional é altamente inclinada ou tem pontos fora da curva.

Figura 8.8 Resumo dos procedimentos de estimação por intervalo de uma média populacional

O desvio padrão σ da população pode ser considerado conhecido?

Sim → Use $\bar{x} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$ — Caso em que σ é conhecido

Não → Use o desvio padrão s da amostra para estimar σ → Use $\bar{x} \pm t_{\alpha/2} \frac{s}{\sqrt{n}}$ — Caso em que σ é desconhecido

NOTAS E COMENTÁRIOS

1. Quando σ é conhecido, a margem de erro, $z_{\alpha/2}(\sigma/\sqrt{n})$, é fixa e é a mesma para todas as amostras de tamanho n. Quando σ é desconhecido, a margem de erro, $t_{\alpha/2}(s/\sqrt{n})$, varia de amostra a amostra. Essa variação ocorre porque o desvio padrão s da amostra varia, dependendo da amostra selecionada. Um valor grande para s produz uma margem de erro maior, ao passo que um valor pequeno para s produz uma margem de erro menor.

2. O que acontece à estimação do intervalo de confiança quando a população é assimétrica? Considere uma população que tem uma inflexão à direita, com grandes valores de dados estendendo a distribuição à direita. Quando existe esse tipo de inflexão, a média amostral $\bar{x}$ e o desvio padrão s da amostra estão positivamente correlacionados. Valores maiores de σ tendem a estar associados a valores maiores de $\bar{x}$. Desse modo, quando $\bar{x}$ é maior que a média populacional, s tende a ser maior que σ. Essa assimetria faz que a margem de erro, $t_{\alpha/2}(s/\sqrt{n})$, seja maior do que seria com σ conhecido. O intervalo de confiança com a margem de erro maior tende a incluir a média populacional μ mais freqüentemente que aquilo que ocorreria se o valor verdadeiro de σ fosse usado. Mas quando $\bar{x}$ é menor que a média populacional, a correlação entre $\bar{x}$ e s faz que a margem de erro seja pequena. Nesse caso, o intervalo de confiança com a margem de erro menor tende a não incluir a média populacional mais freqüentemente que aquilo que ocorreria se soubéssemos o valor de σ e o usássemos. Por esse motivo, recomendamos usar tamanhos de amostra maiores quando se trata de distribuições populacionais altamente assimétricas.

Exercícios

Métodos

11. Para uma distribuição t com 16 graus de liberdade, encontre a área, ou probabilidade, em cada região apresentada a seguir:
    a. À direita de 2,120.
    b. À esquerda de 1,337.
    c. À esquerda de –1,746.
    d. À direita de 2,583.
    e. Entre –2,120 e 2,120.
    f. Entre –1,746 e 1,746.
12. Encontre o(s) valor(es) t em cada um dos seguintes casos:
    a. Área da cauda superior igual a 0,025, com 12 graus de liberdade.
    b. Área da cauda inferior igual a 0,05, com 50 graus de liberdade.
    c. Área da cauda superior igual a 0,01, com 30 graus de liberdade.
    d. Em que 90% da área se situa entre esses dois valores t com 25 graus de liberdade.
    e. Em que 95% da área se situa entre esses dois valores t com 45 graus de liberdade. (Veja na Tabela 2 do Apêndice B uma tabela t mais extensa.)
13. Os dados amostrais seguintes são de uma população normal: 10, 8, 12, 15, 13, 11, 6, 5.

AUTOTESTE

    a. Qual é a estimação por ponto da média populacional?
    b. Qual é a estimação por ponto do desvio padrão da população?
    c. Com 95% de confiança, qual é a margem de erro da estimativa da média populacional?
    d. Qual é o intervalo de confiança de 95% da média populacional?
14. Uma amostra aleatória simples com n =54 produziu uma média amostral igual a 22,5 e um desvio padrão da amostra igual a 4,4. (Veja na Tabela 2 do Apêndice B uma tabela t mais extensa.)
    a. Desenvolva um intervalo de confiança de 90% para a média populacional.
    b. Estabeleça um intervalo de confiança de 95% para a média populacional.
    c. Estipule um intervalo de confiança de 99% para a média populacional.
    d. O que acontece à margem de erro e ao intervalo de confiança quando o grau de confiança é aumentado?

Aplicações

15. A equipe de vendas da Skillings Distributors apresenta semanalmente relatórios que relacionam os contatos feitos com clientes durante a semana. Uma amostra de 65 relatórios semanais exibiu uma média amostral de 19,5 contatos com clientes por semana. O desvio padrão da amostra foi 5,2. Forneça os intervalos de confiança de 90% e 95% correspondentes ao número médio da população de contatos semanais com clientes feitos pela equipe de vendas.

AUTOTESTE

16. O número médio de horas de vôo dos pilotos da Continental Airlines equivale a 49 horas por mês (*The Wall Street Journal*, 25 de fevereiro de 2003). Suponha que essa média tenha se baseado em tempos de vôo reais de uma amostra de 110 pilotos da Continental e que o desvio padrão da amostra tenha sido de 8,5 horas.
    a. Com 95% de confiança, qual é a margem de erro?
    b. Qual é a estimação por intervalo de confiança de 95% do tempo de vôo médio da população de pilotos?
    c. O número médio de horas de vôo dos pilotos da United Airlines equivale a 36 horas por mês. Use os resultados que obteve no item (b) para discutir as diferenças entre os tempos de vôo dos pilotos das duas empresas aéreas. O *Wall Street Journal* publicou que a United Airlines tem o custo de mão-de-obra mais elevado entre todas as empresas aéreas. A informação contida neste exercício oferece subsídios para compreendermos por que a United Airlines poderia esperar custos de mão-de-obra mais elevados?
17. A International Air Transport Association consulta pessoas que viajam a negócios a fim de desenvolver avaliações da qualidade dos aeroportos internacionais. A avaliação máxima possível é 10. Suponha que uma amostra aleatória simples de 50 pessoas que viajam a negócios seja selecionada e

ARQUIVO DA INTERNET
Miami

que cada viajante seja solicitado a fornecer uma avaliação do Aeroporto Internacional de Miami. As avaliações obtidas da amostra de 50 viajantes de negócios são as seguintes:

6	4	6	8	7	7	6	3	3	8	10	4	8
7	8	7	5	9	5	8	4	3	8	5	5	4
4	4	8	4	5	6	2	5	9	9	8	4	8
9	9	5	9	7	8	3	10	8	9	6		

Desenvolva uma estimação por intervalo de confiança de 95% da avaliação média da população para o aeroporto de Miami.

18. Trinta restaurantes de *fast-food*, incluindo o Wendy's, o McDonald's e o Burger King, foram freqüentados durante o verão de 2000 (*The Cincinnatti Enquirer*, 9 de julho de 2000). Durante cada visita, o cliente ia ao *drive-through* e pedia uma refeição básica, por exemplo, uma refeição "*combo*"[4] ou um sanduíche, batatas fritas e um *milk-shake*. Foi registrado o tempo decorrido entre escolher a opção do cardápio e receber o pedido. Os tempos, em minutos, para as 30 visitas foram os seguintes:

ARQUIVO DA INTERNET
Fast Food

0,9	1,0	1,2	2,2	1,9	3,6	2,8	5,2	1,8	2,1
6,8	1,3	3,0	4,5	2,8	2,3	2,7	5,7	4,8	3,5
2,6	3,3	5,0	4,0	7,2	9,1	2,8	3,6	7,3	9,0

a. Apresente uma estimação por ponto da média populacional de tempo gasto nos *drive-throughs* dos restaurantes de *fast-food*.
b. Com 95% de confiança, qual é a margem de erro?
c. Qual é a estimação por intervalo de confiança de 95% para a média populacional?
d. Discuta a assimetria que possa estar presente nessa população. Qual sugestão você apresentaria em uma repetição desse estudo?

19. Uma pesquisa da National Retail Foundation descobriu que as famílias pretendiam gastar uma média de US$ 649 durante o período de festas em dezembro (*The Wall Street Journal*, dezembro de 2002). Suponha que a pesquisa tenha incluído 600 famílias e que o desvio padrão da amostra tenha sido US$ 175.

a. Com 95% de confiança, qual é a margem de erro?
b. Qual é a estimação por intervalo de confiança de 95% para a média populacional?
c. No ano anterior, a média populacional de gastos por família foi de US$ 632. Discuta a mudança nos gastos das festas de fim de ano no período de um ano.

ARQUIVO DA INTERNET
TVtime

20. A American Association of Advertising Agencies publica dados sobre o tempo de propaganda, em minutos, durante meia hora nos programas do horário nobre. Os dados representativos, em minutos, de uma amostra de 20 programas do horário nobre nas principais redes de TV às 8h30 da noite são os seguintes:

6,0	6,6	5,8
7,0	6,3	6,2
7,2	5,7	6,4
7,0	6,5	6,2
6,0	6,5	7,2
7,3	7,6	6,8
6,0	6,2	

Suponha uma população normal e forneça uma estimação por ponto e um intervalo de confiança de 95% referentes ao número médio de minutos de propaganda durante meia hora nos programas de televisão no horário nobre, às 8h30 da noite.

21. As reclamações sobre os preços crescentes dos medicamentos vendidos sob prescrição médica fizeram que o Congresso dos Estados Unidos considerasse leis que obrigassem as empresas de produtos farmacêuticos a oferecer descontos na venda desses medicamentos a idosos que não contassem com os benefícios para aquisição de medicamentos. O *House Government Reform Committee* forneceu

[4] NT: *Combo* – Combinação (de vários itens).

dados sobre o custo de alguns dos medicamentos vendidos com receita mais amplamente usados (*Newsweek*, 8 de maio de 2000). Suponha que os dados apresentados a seguir sejam de uma amostra do custo de prescrição, em dólares, do Zocor, um medicamento usado para reduzir o colesterol.

110 112 115 99 100 98 104 126

Dada uma população normal, qual é a estimação por intervalo de confiança de 95% do custo médio populacional de uma receita médica de Zocor?

22. As primeiras semanas de 2004 foram boas para o mercado de ações. Uma amostra de 25 grandes fundos de capitalização ilimitada (*open-end funds*) apresentou os seguintes retornos no intervalo de um ano, com vencimento em 16 de janeiro de 2004 (*Barron's*, 19 de janeiro de 2004).

7,0	3,2	1,4	5,4	8,5
2,5	2,5	1,9	5,4	1,6
1,0	2,1	8,5	4,3	6,2
1,5	1,2	2,7	3,8	2,0
1,2	2,6	4,0	2,6	0,6

a. Qual é a estimação por ponto do retorno médio populacional no intervalo de um ano, até o presente, para os fundos de capitalização ilimitada?
b. Dado que a população tenha uma distribuição normal, desenvolva um intervalo de confiança de 95% do retorno médio populacional no intervalo de um ano, até o presente, para os fundos de capitalização ilimitada.

8.3 COMO DETERMINAR O TAMANHO DA AMOSTRA

Se uma margem de erro desejada for escolhida antes da amostragem, os procedimentos desta seção poderão ser utilizados para determinar o tamanho de amostra necessário para satisfazer os requisitos da margem de erro.

Ao darmos o conselho prático nas duas seções anteriores, comentamos sobre o papel do tamanho da amostra para produzir bons intervalos de confiança aproximados quando a população não está normalmente distribuída. Nesta seção, concentramo-nos em outro aspecto da questão do tamanho de amostra. Descrevemos como escolher um tamanho de amostra grande o suficiente para produzir uma margem de erro desejada. Para entender como esse processo é feito, retornemos ao caso em que σ é conhecido, apresentado na Seção 8.1. Usando a Equação 8.1, a estimação por intervalo é:

$$\bar{x} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

A quantidade $z_{\alpha/2}(\sigma/\sqrt{n})$ é a margem de erro. Desse modo, observamos que $z_{\alpha/2}$, o desvio padrão σ da população e o tamanho n da amostra se conjugam para determinar a margem de erro.

Assim que escolhermos um coeficiente de confiança, $1-\alpha$, $z_{\alpha/2}$ pode ser determinado. Então, se tivermos um valor para σ, podemos estipular o tamanho n de amostra necessário para fornecer qualquer margem de erro desejada. O desenvolvimento da fórmula utilizada para calcular o tamanho n de amostra necessário é apresentado a seguir.

Digamos que E = a margem de erro desejada:

$$E = z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

A Equação 8.3 pode ser usada para fornecer uma boa recomendação de tamanho de amostra. Entretanto, o julgamento feito pelo analista deve ser usado para determinar se o tamanho de amostra final deve ser ajustado para um valor maior.

Resolvendo para $\sqrt{n}$, temos:

$$\sqrt{n} = \frac{z_{\alpha/2}\sigma}{E}$$

Elevando ao quadrado ambos os termos dessa equação, obtemos a seguinte expressão do tamanho de amostra:

TAMANHO DE AMOSTRA PARA UMA ESTIMAÇÃO POR INTERVALO DE UMA MÉDIA POPULACIONAL

$$n = \frac{(z_{\alpha/2})^2\sigma^2}{E^2} \qquad (8.3)$$

Deve-se especificar um valor planejado para o desvio padrão σ da população a fim de que o tamanho da amostra possa ser determinado. Aqui, discutimos três métodos para se obter o valor planejado de σ.

Esse tamanho de amostra fornece a margem de erro desejada, ao nível de confiança escolhido.

Na Equação 8.3, E é a margem de erro que o usuário está disposto a aceitar e o valor de $z_{\alpha/2}$ decorre diretamente do grau de confiança a ser usado no desenvolvimento da estimação por intervalo. Embora a preferência do usuário deva ser levada em consideração, 95% de confiança é o valor usado com maior freqüência ($z_{0,025}$ =1,96).

Por fim, o uso da Equação 8.3 necessita de um valor para o desvio padrão σ da população. Entretanto, mesmo que σ seja desconhecido, podemos utilizar a Equação 8.3 desde que tenhamos um valor preliminar, ou *valor planejado*, para σ. Na prática, um dos procedimentos seguintes pode ser escolhido.

1. Use a estimativa do desvio padrão da população, calculada a partir de dados de estudos anteriores, como o valor planejado para σ.
2. Use um estudo piloto para selecionar uma amostra preliminar. O desvio padrão amostral da amostra preliminar pode ser usado como o valor planejado para σ.
3. Use o julgamento ou o "melhor palpite" para o valor de σ. Por exemplo, poderíamos começar estimando os maiores e os menores valores de dados da população. A diferença entre os maiores e os menores valores fornece uma estimativa da amplitude dos dados. Finalmente, muitas vezes a amplitude dividida por 4 é sugerida como uma aproximação tosca do desvio padrão e, assim, um valor planejado aceitável para σ.

A Equação 8.3 fornece o tamanho mínimo de amostra necessário para satisfazer os requisitos da margem de erro desejada. Se o tamanho de amostra calculado não for um número inteiro, arredondá-lo para o valor inteiro seguinte produzirá uma margem de erro ligeiramente menor que o necessário.

Vamos demonstrar o uso da Equação 8.3 para determinar o tamanho da amostra, considerando o seguinte exemplo. Um estudo anterior que investigou o custo do aluguel de automóveis nos Estados Unidos revelou que o custo médio para alugar um carro de porte médio era de aproximadamente US$ 55 por dia. Suponha que a organização que realizou esse estudo queira realizar um novo estudo a fim de estimar a média populacional do custo diário de aluguel de automóveis de tamanho médio nos Estados Unidos. Ao projetar o novo estudo, o diretor do projeto especifica que a média populacional do custo de aluguel deve ser estimada com uma margem de erro de US$ 2 e um grau de confiança de 95%.

O diretor do projeto especificou uma margem de erro desejada de E =2, e o grau de confiança de 95% indica $z_{0,025}$ =1,96. Desse modo, precisamos somente de um valor planejado para o desvio padrão σ da população para calcular o tamanho de amostra necessário. Nesse ponto, o analista revisou os dados amostrais do estudo anterior e descobriu que o desvio padrão amostral do custo diário de aluguel era de US$ 9,65. Usando 9,65 como o valor planejado de σ, obtemos:

$$n = \frac{(z_{\alpha/2})^2\sigma^2}{E^2} = \frac{(1,96)^2(9,65)^2}{2^2} = 89,43$$

Assim, o tamanho de amostra do novo estudo precisa ser, no mínimo, de 89,43 aluguéis de automóveis de tamanho médio para satisfazer a exigência de margem de erro igual a US$ 2 determinada pelo diretor do projeto. Nos casos em que o n calculado não for um número inteiro, nós o arredondaremos para o valor inteiro seguinte; portanto, o tamanho de amostra recomendado é de 90 aluguéis de automóveis de tamanho médio.

Exercícios

Métodos

23. Qual tamanho de amostra deve ser selecionado para produzir um intervalo de confiança de 95% com uma margem de erro igual a 10? Suponha que o desvio padrão da população seja 40.

AUTOTESTE

24. Estima-se que a amplitude de um conjunto de dados seja 36.
    a. Qual é o valor planejado do desvio padrão da população?
    b. Com um grau de confiança de 95%, qual tamanho de amostra forneceria uma margem de erro igual a 3?
    c. Com um grau de confiança de 95%, qual tamanho de amostra forneceria uma margem de erro igual a 2?

Aplicações

AUTOTESTE

25. Consulte o exemplo da Scheer Industries na Seção 8.2. Use 6,82 dias como valor planejado para o desvio padrão da população.
    a. Supondo um grau de confiança de 95%, qual tamanho de amostra seria necessário para se obter uma margem de erro de 1,5 dia?
    b. Se a proposição da precisão fosse feita com 90% de confiança, qual tamanho de amostra seria necessário para se obter uma margem de erro de 2 dias?
26. A revista *Bride's* divulgou que o custo médio de um casamento é de US$ 19 mil (*USA Today*, 17 de abril de 2000). Suponha que o desvio médio da população seja US$ 9.400. A *Bride's* planeja usar uma pesquisa anual para monitorar o custo de um casamento. Use 95% de confiança.
    a. Qual é o tamanho de amostra recomendado se a margem de erro desejada for de US$ 1.000?
    b. Qual é o tamanho de amostra recomendado se a margem de erro desejada for de US$ 500?
    c. Qual é o tamanho de amostra recomendado se a margem de erro desejada for de US$ 200?
27. Geralmente se espera que os salários anuais inicias dos diplomados em cursos de pós-graduação em Administração estejam entre US$ 30 mil e US$ 45 mil. Suponha que se deseje uma estimação por intervalo de confiança de 95% da média populacional dos salários anuais iniciais. Qual é o valor planejado para o desvio padrão da população? Qual tamanho de amostra deve ser tomado se a margem de erro desejada for de:
    a. US$ 500?
    b. US$ 200?
    c. US$ 100?
    d. Você recomendaria tentar obter a margem de erro de US$ 100? Explique.
28. A Smith Travel Research fornece informações sobre o custo de pernoites em quartos de hotel em todo o território dos Estados Unidos (*USA Today*, 8 de julho de 2002). Use US$ 2 como a margem de erro desejada e US$ 22,50 como valor planejado para o desvio padrão da população para encontrar o tamanho de amostra recomendado nos itens (a), (b) e (c).
    a. Uma estimação por intervalo de confiança de 90% do custo médio populacional dos quartos de hotel.
    b. Uma estimação por intervalo de confiança de 95% do custo médio populacional dos quartos de hotel.
    c. Uma estimação por intervalo de confiança de 99% do custo médio populacional dos quartos de hotel.
    d. Quando a margem de erro é fixa, o que acontece ao tamanho da amostra quando o grau de confiança é aumentado? Você recomendaria que a Smith Travel Research utilizasse um grau de confiança de 99%? Discuta.
29. O tempo que os habitantes das 15 maiores cidades dos Estados Unidos gastam para ir de casa ao trabalho foi divulgado no *2003 Information Please Almanac*. Suponha que uma amostra aleatória simples preliminar dos habitantes de São Francisco seja usada para desenvolver um valor planejado de 6,25 minutos para o desvio padrão da população.
    a. Se quisermos estimar o tempo médio populacional das viagens ao trabalho para os habitantes de São Francisco com uma margem de erro de 2 minutos, qual tamanho de amostra deve ser usado? Suponha 95% de confiança.
    b. Se quisermos estimar o tempo médio populacional das viagens ao trabalho para os habitantes de São Francisco com uma margem de erro de 1 minuto, qual tamanho de amostra deve ser usado? Suponha 95% de confiança.
30. Durante o primeiro trimestre de 2003, a relação preço/rendimentos (P/R) das ações listadas na Bolsa de Valores de Nova York geralmente variou de 5 a 60 (*The Wall Street Journal*, 7 de março de 2003). Suponha que queiramos estimar a média populacional da relação preço/rendimentos de todas as ações listadas na Bolsa. Quantas ações devem ser incluídas se quisermos uma margem de erro igual a 3? Use 95% de confiança.

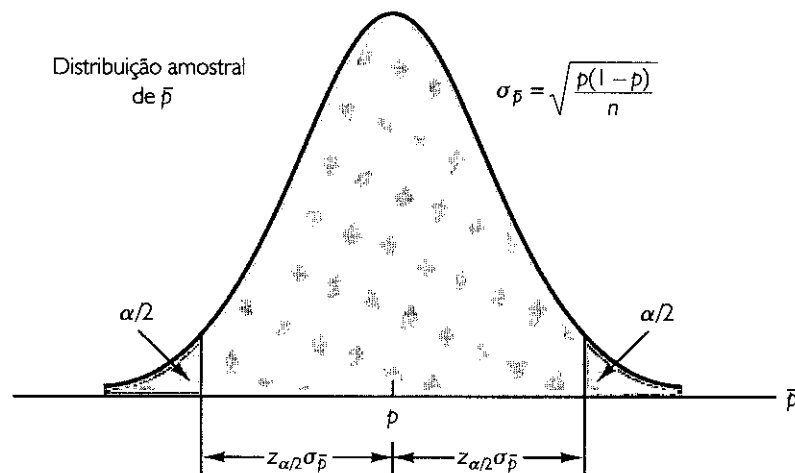
8.4 PROPORÇÃO DA POPULAÇÃO

Na introdução deste capítulo, dissemos que a forma geral de uma estimação por intervalo de uma proporção populacional p é:

$$p \pm \text{Margem de erro}$$

A distribuição amostral de $\bar{p}$ desempenha papel fundamental no cálculo da margem de erro dessa estimação por intervalo.

No Capítulo 7, dissemos que a distribuição amostral de $\bar{p}$ pode ser aproximada por meio de uma distribuição normal quando $np \geq 5$ e $n(1-p) \geq 5$. A Figura 8.9 mostra a aproximação normal da distribuição amostral de $\bar{p}$.

Figura 8.9 Aproximação normal à distribuição amostral de $\bar{p}$

A média da distribuição amostral de $\bar{p}$ é a proporção p da população, e o desvio padrão de $\bar{p}$ é:

$$\sigma_{\bar{p}} = \sqrt{\frac{p(1-p)}{n}} \qquad (8.4)$$

Uma vez que a distribuição amostral de $\bar{p}$ está normalmente distribuída, se escolhermos $z_{\alpha/2}\sigma_{\bar{p}}$ como a margem de erro em uma estimação por intervalo da proporção populacional, saberemos que $100(1-\alpha)\%$ dos intervalos gerados conterão a proporção populacional verdadeira. Mas $\sigma_{\bar{p}}$ não pode ser usado diretamente no cálculo da margem de erro porque p não será conhecido; p é aquilo que estamos tentando estimar. Então, $\bar{p}$ é substituído por p, e a margem de erro de uma estimação por intervalo de uma proporção populacional é dada por:

Quando se desenvolvem intervalos de confiança para proporções, a quantidade $z_{\alpha/2}\sqrt{\bar{p}(1-\bar{p})/n}$ fornece a margem de erro.

$$\text{Margem de erro} = z_{\alpha/2}\sqrt{\frac{\bar{p}(1-\bar{p})}{n}} \qquad (8.5)$$

Com essa margem de erro, a expressão geral da estimação por intervalo de uma proporção populacional é a seguinte:

ESTIMAÇÃO POR INTERVALO DE UMA PROPORÇÃO POPULACIONAL

$$\bar{p} \pm z_{\alpha/2}\sqrt{\frac{\bar{p}(1-\bar{p})}{n}} \qquad (8.6)$$

Em que $1 - \alpha$ é o coeficiente de confiança e $z_{\alpha/2}$ é o valor de z que produz uma área igual a $\alpha/2$ na cauda superior da distribuição normal padrão.

O exemplo a seguir ilustra o cálculo da margem de erro e a estimação por intervalo de uma proporção populacional. Foi realizada uma pesquisa nacional de 900 jogadoras de golfe para saber como as mulheres viam o tratamento que lhes era dado nos cursos de golfe nos Estados Unidos. A pesquisa revelou que 396 das golfistas estavam satisfeitas com a disponibilidade de *tee times*.[5] Desse modo, a estimação por ponto da proporção da população de mulheres golfistas que estão satisfeitas com a disponibilidade de *tee times* é de 396/900 =0,44. Usando a Equação 8.6 e um grau de confiança de 95%,

$$\bar{p} \pm z_{\alpha/2}\sqrt{\frac{\bar{p}(1-\bar{p})}{n}}$$

$$0{,}44 \pm 1{,}96\sqrt{\frac{0{,}44(1-0{,}44)}{900}}$$

$$0{,}44 \pm 0{,}0324$$

Assim, a margem de erro é 0,0324 e a estimação por intervalo de confiança de 95% da proporção populacional é de 0,4076 a 0,4724. Utilizando porcentagens, os resultados da pesquisa nos possibilitam afirmar com 95% de confiança que entre 40,76% e 47,24% de todas as mulheres golfistas estão satisfeitas com a disponibilidade de *tee times*.

Como Determinar o Tamanho da Amostra

Consideremos a questão de qual deve ser o tamanho da amostra para obtermos uma estimativa da proporção populacional a um grau de confiança específico. O fundamento lógico para a determinação do tamanho de amostra para desenvolvermos estimações por intervalo de p é análogo ao fundamento lógico utilizado na Seção 8.3 para estabelecermos o tamanho de amostra para estimar uma média populacional.

Anteriormente, nesta seção, dissemos que a margem de erro associada a uma estimação por intervalo de uma proporção populacional é $z_{\alpha/2}\sqrt{\bar{p}(1-\bar{p})/n}$. A margem de erro baseia-se no valor de $z_{\alpha/2}$, na proporção $\bar{p}$ da amostra e no tamanho n da amostra. Tamanhos de amostra maiores produzem uma margem de erro menor e uma precisão melhor.

Digamos que E denote a margem de erro desejada.

$$E = z_{\alpha/2}\sqrt{\frac{\bar{p}(1-\bar{p})}{n}}$$

Isolar n nessa equação produz uma fórmula do tamanho de amostra que fornecerá uma margem de erro de tamanho E.

$$n = \frac{(z_{\alpha/2})^2\bar{p}(1-\bar{p})}{E^2}$$

Observe, entretanto, que não podemos usar essa fórmula para calcular o tamanho de amostra que produzirá a margem de erro desejada, porque $\bar{p}$ somente será conhecido depois de selecionarmos a amostra. O que precisamos, então, é de um valor planejado para $\bar{p}$ que possa ser usado para fazermos o cálculo. Usando p^* para denotar o valor planejado de $\bar{p}$, podemos utilizar a fórmula apresentada a seguir para calcular o tamanho de amostra que produzirá uma margem de erro de tamanho E.

[5] NT: *Tee time* ("hora de saída") – Momento em que há um *tee* (ponto a partir do qual se bate a primeira tacada em cada buraco) disponível (Golfe).

TAMANHO DA AMOSTRA PARA UMA ESTIMAÇÃO POR INTERVALO DE UMA PROPORÇÃO POPULACIONAL

$$n = \frac{(z_{\alpha/2})^2 p^*(1 - p^*)}{E^2} \tag{8.7}$$

Na prática, o valor planejado p^* pode ser escolhido por meio de um dos seguintes procedimentos.

1. Use a proporção amostral de uma amostra anterior das mesmas unidades ou de unidades similares.
2. Use um estudo piloto para selecionar uma amostra preliminar. A proporção amostral dessa amostra pode ser usada como o valor planejado, p^*.
3. Use o julgamento ou o "melhor palpite" para o valor de p^*.
4. Se nenhuma das alternativas anteriores for apropriada, use o valor planejado de $p^* = 50$.

Retornemos à pesquisa das mulheres golfistas e vamos supor que a empresa esteja interessada em realizar uma nova pesquisa para estimar a proporção atual da população de mulheres praticantes do golfe que estão satisfeitas com a disponibilidade de *tee times*. Qual deve ser o tamanho da amostra se o diretor da pesquisa quiser estimar a proporção populacional com uma margem de erro de 0,025, com 95% de confiança? Com $E = 0{,}025$ e $z_{\alpha/2} = 1{,}96$, precisamos de um valor planejado p^* para responder à questão do tamanho da amostra. Utilizando o resultado da pesquisa anterior, em que $\bar{p} = 0{,}44$ como o valor planejado p^*, a Equação 8.7 mostra que:

$$n = \frac{(z_{\alpha/2})^2 p^*(1 - p^*)}{E^2} = \frac{(1{,}96)^2(0{,}44)(1 - 0{,}44)}{(0{,}025)^2} = 1.514{,}5$$

Tabela 8.5 Alguns valores possíveis para $p^*(1 - p^*)$

p^*	$p^*(1 - p^*)$	
0,10	(0,10)(0,90) = 0,09	
0,30	(0,30)(0,70) = 0,21	
0,40	(0,40)(0,60) = 0,24	
0,50	(0,50)(0,50) = 0,25	← O maior valor para $p^*(1 - p^*)$
0,60	(0,60)(0,40) = 0,24	
0,70	(0,70)(0,30) = 0,21	
0,90	(0,90)(0,10) = 0,09	

Desse modo, o tamanho da amostra deve ter, no mínimo, 1.514,5 mulheres golfistas para que o requisito de margem de erro seja satisfeito. O arredondamento para o valor inteiro seguinte indica que uma amostra de 1.515 mulheres golfistas é recomendada para que o requisito de margem de erro seja cumprido.

A quarta alternativa sugerida para se escolher um valor planejado de p^* é usar $p^* = 0{,}50$. Esse valor de p^* freqüentemente é usado quando não há nenhuma outra informação disponível. Para entender o porquê, observe que o numerador da Equação 8.7 mostra que o tamanho de amostra é proporcional à quantidade $p^*(1 - p^*)$. Um valor maior para a quantidade $p^*(1 - p^*)$ resultará em um tamanho de amostra maior. A Tabela 8.5 apresenta alguns valores possíveis para $p^*(1 - p^*)$. Note que o maior valor de $p^*(1 - p^*)$ ocorre quando $p^* = 0{,}50$. Assim, no caso de qualquer incerteza a respeito de um valor planejado apropriado, sabemos que $p^* = 0{,}50$ apresentará a recomendação do maior tamanho de amostra. De fato, sentimo-nos seguros em recomendar o maior tamanho de amostra possível. Se a proporção amostral vier a ser diferente do valor planejado de 0,50, a margem de erro será menor que o previsto. Logo, ao usar $p^* = 0{,}50$, garantimos que o tamanho da amostra será suficiente para obtermos a margem de erro desejada.

No exemplo das mulheres golfistas, um valor planejado de $p^* = 0{,}50$ teria produzido o seguinte tamanho de amostra:

$$n = \frac{(z_{\alpha/2})^2 p^*(1 - p^*)}{E^2} = \frac{(1{,}96)^2(0{,}50)(1 - 0{,}50)}{(0{,}025)^2} = 1.536{,}6$$

Desse modo, um tamanho de amostra ligeiramente maior, de 1.537 mulheres golfistas, seria recomendado.

NOTAS E COMENTÁRIOS

A margem de erro desejada para estimar uma proporção populacional é quase sempre 0,10 ou menos. Em pesquisas de opinião pública realizadas por organizações como o Instituto Gallup e a Harris,[6] uma margem de erro de 0,03 ou 0,04 é comum. Com essas margens de erro, a Equação 8.7 quase sempre fornecerá um tamanho de amostra que é grande o bastante para satisfazer à condição essencial de $np \geq 5$ e $n(1-p) \geq 5$ para que se possa usar uma distribuição normal como uma aproximação à distribuição amostral de $\bar{x}$.

Exercícios

Métodos

AUTOTESTE

31. Uma amostra aleatória simples de 400 pessoas apresentou 100 respostas Sim.
    a. Qual é a estimação por ponto da proporção da população que apresentaria respostas Sim?
    b. Qual é sua estimativa do erro padrão da população, $\sigma_{\bar{p}}$?
    c. Calcule o intervalo de confiança de 95% para a proporção populacional?
32. Uma amostra aleatória simples de 800 elementos gera uma proporção amostral $\bar{p} = 0{,}70$.
    a. Forneça um intervalo de confiança de 90% para a proporção populacional.
    b. Providencie um intervalo de confiança de 95% para a proporção populacional.
33. Em uma pesquisa, o valor planejado da proporção populacional é $p^* = 0{,}35$. Qual tamanho de amostra deve ser tomado para produzir um intervalo de confiança de 95% com uma margem de erro de 0,05?
34. Com 95% de confiança, qual tamanho de amostra deve ser tomado para se obter uma margem de erro de 0,03 para a estimativa de uma proporção populacional? Suponha que não haja dados históricos disponíveis para que se possa desenvolver um valor planejado para p^*.

Aplicações

AUTOTESTE

35. Uma pesquisa de 611 funcionários de escritório investigou seus hábitos de atendimento ao telefone, incluindo a freqüência com que cada funcionário de escritório era capaz de atender às chamadas telefônicas e com qual freqüência as chamadas telefônicas chegavam diretamente ao correio de voz (*USA Today*, 21 de abril de 2002). Ao todo, 281 funcionários de escritório indicaram que nunca precisavam do correio de voz e que eram capazes de responder a cada chamada telefônica.
    a. Qual é a estimação por ponto da proporção da população de funcionários de escritório que são capazes de atender a cada chamada telefônica?
    b. Com 90% de confiança, qual é a margem de erro?
    c. Qual é o intervalo de confiança de 90% da proporção da população de funcionários de escritório que são capazes de atender a cada chamada telefônica?
36. Uma pesquisa realizada pela Society for Human Resource Management perguntou a 346 pessoas que procuravam emprego por que os empregados trocam de emprego tão freqüentemente (*The Wall Street Journal*, 28 de março de 2000). A resposta mais escolhida (152 vezes) foi "melhor remuneração em outro lugar".
    a. Qual é a estimação por ponto da proporção de pessoas que procuram emprego que escolheriam "melhor remuneração em outro lugar" como a razão para trocar de emprego?
    b. Qual é a estimação por intervalo de confiança de 95% da proporção populacional?
37. A Towers Perrin, uma firma de consultoria em recursos humanos de Nova York, realizou uma pesquisa de 1.100 empregados de empresas de médio e grande portes para determinar qual seria o nível de insatisfação dos empregados com seus empregos (*The Wall Street Journal*, 29 de janeiro de 2003). Ao todo, 473 empregados indicaram que estavam fortemente insatisfeitos com suas experiências de trabalho atuais.

[6] NT: Louis Harris & Associates.

a. Qual é a estimação por ponto da proporção da população de empregados que estão fortemente insatisfeitos com suas experiências de trabalho atuais?
b. Com 95% de confiança, qual é a margem de erro?
c. Qual é o intervalo de confiança de 95% da proporção populacional de empregados que estão fortemente insatisfeitos com suas experiências de trabalho atuais?
d. A Towers Perrin estima que custa aos empregadores 1/3 do salário anual de um empregado que trabalha por hora para encontrar um substituto, e até 1,5 vezes o salário anual para encontrar um substituto para um empregado que recebe altos salários. Qual mensagem essa pesquisa transmite aos empregadores?

38. Dados sobre o perfil do público coletados no site da ESPN SportsZone mostraram que 26% dos usuários eram mulheres (*USA Today*, 21 de janeiro de 1998). Suponha que essa porcentagem tenha se baseado em uma amostra de 400 usuários.
a. Com 95% de confiança, qual é a margem de erro associada à proporção estimada de usuários que são mulheres?
b. Qual é o intervalo de 95% de confiança relativo à proporção populacional de usuários do site da ESPN SportsZone que são mulheres?
c. Qual tamanho de amostra deve ser tomado se a margem de erro desejada for igual a 0,03?

AUTOTESTE

39. Uma pesquisa realizada pelo *Employee Benefit Research Institute* explorou as razões pelas quais os empregadores dos pequenos negócios oferecem um plano de aposentadoria aos seus empregados (*USA Today*, 4 de abril de 2000). A razão "vantagem competitiva no recrutamento e retenção de funcionários" foi antecipada 33% das vezes.
a. Qual tamanho de amostra deve ser recomendado se a meta de uma pesquisa for estimar a proporção de empregadores de pequenos negócios que oferecem um plano de aposentadoria principalmente em função da "vantagem competitiva no recrutamento e retenção de funcionários", com uma margem de erro de 0,03? Use um grau de confiança de 95%.
b. Repita o item (a) utilizando 99% de confiança.

40. O recorde de 61 *home runs*[7] do beisebol profissional em uma temporada foi mantido durante 37 anos por Roger Maris, dos New York Yankees. Entretanto, entre 1998 e 2001, três jogadores – Mark McGwire, Sammy Sosa e Harry Bonds – quebraram as marcas obtidas por Maris, e Bonds mantém o recorde atual de 73 *home runs* em uma única temporada. Considerando a quebra do recorde de *home runs* mantido durante tanto tempo, e com muitos outros recordes absurdos sendo fixados, surgiu a suspeita de que os jogadores de beisebol poderiam estar usando esteróides – as drogas ilegais para aumentar a musculatura. Uma pesquisa de opinião promovida conjuntamente pelo jornal *USA Today*, CNN e Instituto Gallup revelou que 86% dos torcedores de beisebol acham que os jogadores profissionais de beisebol deveriam ser submetidos a testes de detecção de esteróides (*USA Today*, 8 de julho de 2002). Se 650 torcedores de beisebol fossem incluídos na amostra, calcule qual seria a margem de erro e o intervalo de confiança de 95% da proporção populacional de torcedores de beisebol que acham que os jogadores profissionais de beisebol deveriam ser submetidos a testes de detecção de esteróides.

41. Uma pesquisa do comércio varejista realizada pela American Express revelou que 16% dos consumidores norte-americanos usaram a internet para comprar presentes nas festas de fim de ano (*USA Today*, 18 de janeiro de 2000). Se 1.285 consumidores tiverem participado da pesquisa, qual é a margem de erro e qual é a estimação por intervalo da proporção populacional de consumidores que usam a internet para comprar presentes? Use 95% de confiança.

42. Uma pesquisa realizada conjuntamente pelo jornal *USA Today*, CNN e Instituto Gallup para a campanha à Presidência da República tomou como amostra 491 eleitores em potencial em junho (*USA Today*, 9 de junho de 2000). Uma das principais finalidades da pesquisa era obter uma estimativa da proporção dos eleitores em potencial que eram favoráveis a cada candidato. Suponha um valor planejado de $p^* = 0{,}50$ e um grau de confiança de 95%.

---

[7] NT: *Home run* – Jogada máxima de ataque, em que a bola é rebatida para fora do campo de jogo e permite ao rebatedor percorrer todas as bases e marcar um *run* (pontuação por percorrer de maneira bem-sucedida todas as quatro bases) (Beisebol).

a. Para $p^* = 0{,}50$, qual foi a margem de erro planejada para a pesquisa realizada em junho?

b. Quanto mais se aproximam as eleições de novembro, maior precisão e menores margens de erro são desejadas. Suponha que as seguintes margens de erro sejam solicitadas para as pesquisas a serem realizadas durante a campanha à Presidência da República. Calcule o tamanho de amostra recomendado para cada pesquisa.

Pesquisa	Margem de Erro
Setembro	0,04
Outubro	0,03
Início de novembro	0,02
Véspera das eleições	0,01

43. Uma pesquisa realizada pela Phoenix Wealth Management/Harris Interactive de 1.500 indivíduos que possuem riqueza líquida de US$ 1 milhão ou mais forneceu uma série de estatísticas sobre as pessoas ricas (*Business Week*, 22 de setembro de 2003). Os três anos anteriores foram ruins para o mercado de ações, e isso motivou algumas das perguntas que foram feitas.

a. A pesquisa revelou que 53% dos entrevistados perderam 25% ou mais do valor de suas carteiras de ações ao longo dos últimos três anos. Desenvolva um intervalo de confiança de 95% da proporção de pessoas ricas que perderam 25% ou mais do valor de suas carteiras de ações ao longo dos últimos três anos.

b. A pesquisa revelou que 31% dos entrevistados achavam que precisavam poupar mais para a aposentadoria, para compensar aquilo que haviam perdido. Desenvolva um intervalo de confiança de 95% relativo à proporção populacional.

c. Cinco por cento dos entrevistados doaram US$ 25 mil ou mais para obras assistenciais ao longo do ano anterior. Desenvolva um intervalo de confiança de 95% relativo à proporção de quem doou US$ 25 mil ou mais para obras assistenciais.

d. Compare a margem de erro das estimações por intervalo dos itens (a), (b) e (c). Como a margem de erro está relacionada a $\bar{p}$? Quando a mesma amostra é usada para estimar uma série de proporções, qual das proporções deve ser usada para se escolher o valor planejado p^*? Por que você acha que $p^* = 0{,}50$ freqüentemente é usado nesses casos?

Resumo

Neste capítulo, apresentamos métodos para o desenvolvimento de estimações por intervalo de uma média da população e de uma proporção da população. Um estimador por ponto pode produzir ou não produzir uma boa estimativa de um parâmetro populacional. O uso de uma estimação por intervalo fornece uma medida da precisão de uma estimativa.

Tanto a estimação por intervalo da média da população como a proporção populacional têm a seguinte forma: estimação por ponto ± margem de erro.

Apresentamos as estimações por intervalo de uma média populacional relativas a dois casos. No caso em que σ é conhecido, dados históricos ou outras informações são utilizados para desenvolver uma estimativa de σ antes de se extrair a amostra. Então, a análise dos novos dados amostrais é realizada baseando-se no pressuposto de que σ é conhecido. No caso em que σ é desconhecido, os dados amostrais são utilizados para estimar tanto a média populacional como o desvio padrão da população. A escolha final de qual procedimento de estimação por intervalo se deve usar depende do entendimento do analista a respeito de qual método produz a melhor estimativa de σ.

No caso em que s é conhecido, o procedimento de estimação por intervalo baseia-se no valor pressuposto de σ e no uso da distribuição normal padrão. No caso em que σ é desconhecido, o procedimento de estimação por intervalo usa o desvio padrão s da amostra e a distribuição t. Em ambos os casos, a qualidade das estimações por intervalo obtidas depende da distribuição da população e do tamanho da amostra. Se a população estiver normalmente distribuída, as estimações por intervalo serão exatas em ambos os casos, até mesmo para tamanhos de amostra pequenos. Se a população não estiver normalmente distribuída, as estimações por intervalo serão aproximadas. Tamanhos de amostra maiores produzirão melhores aproximações, mas, quanto mais assimétrica for a população, maior deve ser o tamanho da amostra para se obter uma boa aproximação. Um conselho prático a respeito do tamanho de amostra necessário para se obter uma boa aproximação foi incluído nas Seções 8.1 e 8.2. Na maioria dos casos, um tamanho de amostra igual a 30 ou mais produzirá bons intervalos de confiança aproximados.

A forma geral da estimação por intervalo de uma proporção populacional é $\bar{p}$ ± margem de erro. Na prática, os tamanhos de amostra usados nas estimações por intervalo de uma proporção populacional geralmente são grandes. Desse modo, um procedimento de estimação por intervalo baseia-se na distribuição normal padrão.

Uma margem de erro desejada, muitas vezes, é especificada antes de se desenvolver um plano de amostragem. Mostramos como escolher um tamanho de amostra grande o bastante para produzir a precisão desejada.

Glossário

Estimativa do intervalo Uma estimativa de um parâmetro da população que fornece um intervalo no qual se acredita que está o valor do parâmetro. Em relação às estimações por intervalo deste capítulo, ele tem a forma: estimação por ponto ± margem de erro.

Margem de erro O valor ± que é adicionado e subtraído de uma estimação por ponto a fim de se desenvolver uma estimação por intervalo de um parâmetro populacional.

σ conhecido O caso em que dados históricos ou outras informações produzem um bom valor para o desvio padrão da população antes de se tomar a amostra. O procedimento de estimação por intervalo usa esse valor conhecido de σ para calcular a margem de erro.

σ desconhecido O caso mais comum, em que não existe nenhuma base boa, para se estimar o desvio padrão da população antes de se tomar a amostra. O procedimento de estimação por intervalo usa o desvio padrão σ da amostra para calcular a margem de erro.

Grau de confiança A confiança associada a uma estimação por intervalo. Por exemplo, se um procedimento de estimação por intervalo produz intervalos de maneira que 95% deles incluem o parâmetro populacional, diz-se que a estimação por intervalo foi construída com um grau de confiança de 95%.

Coeficiente de confiança O grau de confiança expresso como um valor decimal. Por exemplo, 0,95 é o coeficiente de confiança de um grau de confiança de 95%.

Intervalo de confiança Outro nome para estimação por intervalo.

Distribuição *t* Uma família de distribuições probabilísticas que podem ser usadas para desenvolver uma estimação por intervalo de uma média populacional quando quer que o desvio padrão σ da população seja desconhecido e seja estimado pelo desvio padrão s da amostra.

Graus de liberdade Um parâmetro da distribuição *t*. Quando a distribuição *t* é usada no cálculo de uma estimação por intervalo de uma média populacional, a distribuição *t* apropriada tem $n - 1$ graus de liberdade, em que n é o tamanho da amostra aleatória simples.

Fórmulas-Chave

Estimação por Intervalo de uma Média da População: Caso em que σ é Conhecido

$$\bar{x} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \tag{8.1}$$

Estimação por Intervalo de uma Média da População: Caso em que σ é Desconhecido

$$\bar{x} \pm t_{\alpha/2}\frac{s}{\sqrt{n}} \tag{8.2}$$

Tamanho da Amostra para uma Estimação por Intervalo de uma Média da População

$$n = \frac{(z_{\alpha/2})^2\sigma^2}{E^2} \tag{8.3}$$

Estimação por Intervalo de uma Proporção da População

$$\bar{p} \pm z_{\alpha/2}\sqrt{\frac{\bar{p}(1-\bar{p})}{n}} \tag{8.6}$$

Tamanho da Amostra para uma Estimação por Intervalo de uma Proporção da População

$$n = \frac{(z_{\alpha/2})^2 p^*(1-p^*)}{E^2} \tag{8.7}$$

Exercícios Suplementares

44. Uma pesquisa feita com pessoas que compram sua primeira casa revelou que a média da renda familiar anual era de US$ 50 mil (CNBC.com, 11 de julho de 2000). Suponha que a pesquisa tenha usado uma amostra de 400 pessoas e suponha que o desvio padrão da população seja US$ 20.500.
    a. Com 95% de confiança, qual é a margem de erro desse estudo?
    b. Qual é o intervalo de confiança de 95% relativo à média da renda familiar anual da população de pessoas que compram sua primeira casa?
45. Uma pesquisa realizada pela American Automobile Association mostrou que uma família de quatro pessoas gasta em média US$ 215,60 por dia enquanto está em férias. Suponha que uma amostra de 64 famílias de quatro pessoas que tenham ido passar as férias em Niagara Falls resultasse em uma média amostral de US$ 252,45 por dia e um desvio padrão amostral de US$ 74,50.
    a. Desenvolva uma estimação por intervalo de confiança de 95% da quantia média gasta por dia por uma família de quatro pessoas que visita Niagara Falls.
    b. Com base no intervalo de confiança do item (a), parece que a quantia média populacional gasta por dia pelas famílias que visitam Niagara Falls difere da média registrada pela American Automobile Association. Explique.
46. O filme *Harry Potter e a Pedra Filosofal* quebrou, em seu lançamento, o recorde de bilheteria anteriormente mantido pelo filme *O Mundo Perdido – Jurassic Park* (*The Wall Street Journal*, 19 de novembro de 2001). Uma amostra de 100 cinemas revelou que a média de renda bruta em três dias do fim de semana foi de US$ 25.467 por cinema. O desvio padrão da amostra foi de US$ 4.980.
    a. Qual é a margem de erro desse estudo? Use 95% de confiança.
    b. Qual é a estimação por intervalo de confiança de 95% relativa à média populacional de renda bruta por cinema nos fins de semana?
    c. O filme *O Mundo Perdido – Jurassic Park* arrecadou US$ 72,1 milhões em três dias no seu primeiro fim de semana. *Harry Potter e a Pedra Filosofal* foi apresentado em 3.672 cinemas. Qual é a estimativa do total que *Harry Potter e a Pedra Filosofal* arrecadou em três dias no seu primeiro fim de semana?
    d. Um artigo da Associated Press afirmou que *Harry Potter* "estraçalhou", em seu lançamento, o recorde de bilheteria que era mantido pelo filme *O Mundo Perdido – Jurassic Park.* Os resultados que você obteve concordam com essa afirmação?
47. Muitos observadores do mercado de valores dizem que quando a relação preço/rendimentos (P/R) das ações ultrapassa 20, o mercado está "superavaliado". A relação preço/rendimentos é o preço das ações dividido pelo rendimento obtido nos 12 meses mais recentes. Suponha que você esteja interessado em verificar se o mercado está superavaliado, e que também gostaria de saber qual proporção de empresas paga dividendos. Uma amostra aleatória de 30 empresas listadas na Bolsa de Valores de Nova York (Nyse) é apresentada (*Barron's*, 19 de janeiro de 2004).

Empresa	Dividendo	Relação Preço/ Rendimentos	Empresa	Dividendo	Relação Preço/ Rendimentos
Albertsons	Sim	14	NY Times A	Sim	25
BRE Prop	Sim	18	Omnicare	Sim	25
CityNtl	Sim	16	PallCp	Sim	23
DelMonte	Não	21	PubSvcEnt	Sim	11
EnrgzHldg	Não	20	SensientTch	Sim	11
Ford Motor	Sim	22	SmtProp	Sim	12
Gildan A	No	12	TJX Cos	Sim	21
HudsnUtdBcp	Sim	13	Thomson	Sim	30
IBM	Sim	22	USB Hldg	Sim	12
JeffPilot	Sim	16	US Restr	Sim	26
KingswayFin	Não	6	Varian Med	Não	41
Libbey	Sim	13	Visx	Não	72
MasoniteIntl	Não	15	Waste Mgt	Não	23
Motorola	Sim	68	Wiley A	Sim	21
Ntl City	Sim	10	Yum Brands	Não	18

ARQUIVO DA INTERNET
NYSEStocks

a. Qual é a estimação por ponto da relação preço/rendimentos (P/R) das ações listadas na Bolsa de Valores de Nova York (Nyse)? Desenvolva um intervalo de confiança de 95%.

b. Com base em sua resposta ao item (a), você acredita que o mercado está superavaliado?

c. Qual é a estimação por ponto da proporção de empresas listadas na Nyse que pagam dividendos? O tamanho da amostra é suficientemente grande para justificar o uso da distribuição normal para construir um intervalo de confiança para essa proporção? Por quê?

ARQUIVO DA INTERNET
Flights

48. A US Airways realizou uma série de estudos que indicaram que poderiam obter uma economia substancial se estimulassem os clientes participantes do programa de milhagem (*frequent flyer*) Dividend Miles a resgatar as milhas ganhas e a reservar seus vôos-prêmio pelo sistema on-line (*US Airways Attache*, fevereiro de 2003). Um estudo coletou dados sobre a quantidade de tempo necessário para que as pessoas resgatem as milhas ganhas e reservem os vôos recebidos como prêmio pelo telefone. Uma amostra apresentando o tempo em minutos correspondente a 150 reservas de vôos-prêmio pelo telefone está contida no conjunto de dados (*data set*) intitulado Flights. Use o Minitab ou o Excel para auxiliá-lo a responder às seguintes questões:

a. Qual é a média amostral do número de minutos necessários para reservar um vôo-prêmio pelo telefone?

b. Qual é o intervalo de confiança de 95% da média populacional correspondente ao tempo necessário para reservar um vôo-prêmio pelo telefone?

c. Suponha que um(a) atendente trabalhe 7,5 horas por dia atendendo ao telefone. Quantos vôos-prêmio um(a) atendente é capaz de manipular em um dia?

d. Discuta as razões pelas quais essa informação deu suporte aos planos da US Airways para utilizar um sistema on-line para reduzir os custos.

ARQUIVO DA INTERNET
ActTemps

49. Uma pesquisa feita pela Accountemps pediu a uma amostra de 200 executivos que fornecessem dados sobre o número de minutos por dia que os funcionários de escritório gastavam para localizar itens mal rotulados, mal arquivados ou colocados fora do lugar. Dados coerentes com essa pesquisa estão contidos no conjunto de dados ActTemps.

a. Use o arquivo ActTemps para desenvolver uma estimação por ponto do número de minutos por dia que os funcionários de escritório gastam para localizar itens mal rotulados, mal arquivados ou colocados fora do lugar.

b. Qual é o desvio padrão da amostra?

c. Qual é o intervalo de confiança de 95% da média do número de minutos gastos por dia?

50. Foram realizados testes do consumo de combustível de determinado modelo de automóvel. Se for desejado um intervalo de confiança de 98%, com uma margem de erro de 1 milha (1.609 m) por galão (3,78 litros), quantos automóveis deveriam ser usados no teste? Suponha que os testes preliminares de consumo de combustível indiquem que o desvio padrão é de 2,6 milhas (4,18 km) por galão.

51. Para desenvolver a programação de horários de consulta médica, um centro médico quer uma estimativa do tempo médio que um membro da equipe gasta para atender a cada paciente. Qual tamanho de amostra deve ser tomado se a margem de erro desejada é de dois minutos, com um grau de confiança de 95%? Use um valor planejado de oito minutos para o desvio padrão da população.

52. Dados dos salários anuais mais bonificações recebidos pelos CEOs das empresas são publicados na Annual Pay Survey (Pesquisa de Salários Anuais) da revista *Business Week*. Uma amostra preliminar revelou que o desvio padrão é igual a US$ 675, sendo os dados fornecidos em milhares de dólares. Quantos CEOs devem estar contidos em uma amostra se quisermos obter uma estimativa da média populacional dos salários anuais mais bonificações, com uma margem de erro de US$ 100 mil? (*Nota*: A margem de erro desejada seria $E = 100$ se os dados forem expressos em milhares de dólares.) Use 95% de confiança.

53. O National Center for Education Statistics divulgou que 47% dos estudantes universitários trabalham para pagar os gastos de estudo e moradia. Suponha que uma amostra de 450 estudantes tenha sido usada nesse estudo.

a. Forneça um intervalo de confiança de 95% relativo à proporção populacional de estudantes universitários que trabalham para pagar os gastos de estudo e moradia.

b. Providencie um intervalo de confiança de 99% relativo à proporção populacional de estudantes universitários que trabalham para cobrir os gastos de estudo e moradia.

c. O que acontece à margem de erro quando o intervalo de confiança de 95% é aumentado para 99%?

54. Uma pesquisa do jornal *USA Today*, CNN e Instituto Gallup realizada com 369 pais trabalhadores revelou que 200 deles disseram dedicar pouquíssimo tempo aos filhos em razão dos compromissos de trabalho.
    a. Qual é a estimação por ponto da proporção da população de pais trabalhadores que afirmam dedicar pouco tempo aos filhos em virtude dos compromissos de trabalho?
    b. Com 95% de confiança, qual é a margem de erro?
    c. Qual é a estimação por intervalo de confiança da proporção populacional de pais trabalhadores afirmam dedicar pouco tempo aos filhos em conseqüência dos compromissos de trabalho?
55. Qual desses itens você teria mais dificuldade para abrir mão: seu computador ou sua televisão? Em uma pesquisa recente com 1.677 usuários de internet norte-americanos, 74% dos jovens da elite tecnológica (média de idade, 22 anos) dizem que seria muito difícil abrir mão do computador (*PC Magazine*, 3 de fevereiro de 2004). Somente 48% deles dizem que seria muito difícil desistir da televisão.
    a. Desenvolva um intervalo de confiança de 95% relativo à proporção dos jovens da elite tecnológica que achariam muito difícil abrir mão do computador.
    b. Estabeleça um intervalo de confiança de 99% relativo à proporção dos jovens da elite tecnológica que achariam muito difícil abrir mão do computador.
    c. Em qual dos casos, item (a) ou item (b), a margem de erro é maior? Explique o porquê.
56. Uma pesquisa feita pela Roper Starch perguntou a empregados com idades de 18 a 29 anos se eles prefeririam um seguro-saúde melhor ou um aumento de salário (*USA Today*, 5 de setembro de 2000). Responda às perguntas a seguir supondo que 340 de 500 empregados disseram que prefeririam um seguro-saúde melhor em vez de um aumento.
    a. Qual é a estimação por ponto da proporção de empregados com idades de 18 a 29 anos que prefeririam um seguro-saúde melhor?
    b. Qual é a estimação por intervalo de confiança da proporção populacional?
57. O *2003 Statistical Abstract of the United States* divulgou a porcentagem de pessoas com idades a partir de 18 anos que fumam. Suponha que um estudo idealizado para coletar dados dos fumantes e não-fumantes utilize uma estimativa preliminar de 0,30 correspondente à proporção dos fumantes.
    a. Qual tamanho de amostra deve ser tomado para estimar a proporção dos fumantes na população, com uma margem de erro de 0,02? Use 95% de confiança.
    b. Suponha que o estudo utilize sua recomendação de tamanho de amostra do item (a) e encontre 520 fumantes. Qual é a estimação por ponto da proporção de fumantes na população?
    c. Qual é o intervalo de confiança da proporção de fumantes na população?
58. Uma famosa empresa de cartões de crédito deseja estimar a proporção dos portadores de cartão de crédito que apresentam um saldo diferente de zero no fim do mês e incorrem na cobrança de juros. Suponha que a margem de erro desejada seja de 0,03, com 98% de confiança.
    a. Qual tamanho de amostra deve ser selecionado considerando-se que há a previsão de que aproximadamente 70% dos portadores de cartão de crédito da empresa mantêm saldos diferentes de zero no fim do mês?
    b. Qual tamanho de amostra deveria ser selecionado se nenhum valor planejado para a proporção pudesse ser especificado?
59. Em uma pesquisa, 200 pessoas foram solicitadas a identificar suas principais fontes de notícias; 110 declararam que a principal fonte de informação eram os noticiários de televisão.
    a. Construa um intervalo de confiança de 95% relativo à proporção de pessoas da população que consideram a televisão sua principal fonte de notícias.
    b. Qual tamanho de amostra seria necessário para estimar a proporção populacional com uma margem de erro de 0,05, com 95% de confiança?
60. Não obstante os horários das empresas aéreas e o custo serem fatores importantes para as pessoas que viajam a negócios ao escolherem uma empresa aérea, uma pesquisa realizada pelo jornal *USA Today* revelou que as pessoas que fazem viagens de negócios mencionam o programa de milhagem (*frequent flyer*) como o fator mais importante. De uma amostra de $n = 1.993$ viajantes de negócios que responderam à pesquisa, 618 mencionaram um programa de milhagem como o fator mais importante.
    a. Qual é a estimação por ponto da proporção da população de pessoas que fazem viagens de negócios que acreditam que um programa de milhagem é o fator mais importante ao escolherem uma empresa aérea?

b. Desenvolva uma estimação por intervalo de confiança da proporção populacional.
c. Qual tamanho de amostra seria necessário para registrar a margem de erro de 0,01, com 95% de confiança? Você recomendaria ao *USA Today* tentar fornecer esse grau de precisão? Por quê?

Estudo de Caso 1 – Bock Investment Services

A meta da Bock Investment Services (BIS) é tornar-se a líder em serviços de consultoria do mercado financeiro na Carolina do Sul. Para oferecer melhores serviços aos seus clientes atuais e para atrair novos clientes, a BIS desenvolveu um boletim informativo semanal. A BIS está considerando a possibilidade de adicionar um novo destaque (*feature*) ao boletim informativo, o qual relate os resultados de uma pesquisa telefônica semanal de gerentes de fundos financeiros. Para investigar a viabilidade de oferecer esse serviço, e para determinar qual tipo de informação incluir no boletim, a BIS selecionou uma amostra aleatória simples de 45 fundos do mercado financeiro. Uma parte dos dados obtidos é apresentada na Tabela 8.6, que registra os ativos e os rendimentos dos fundos financeiros no último período de sete a 30 dias. Antes de telefonar aos gerentes de fundos do mercado financeiro para obter dados adicionais, a BIS decidiu realizar algumas análises preliminares dos dados já coletados.

Relatório Administrativo

1. Use a estatística descritiva apropriada para sintetizar os dados sobre ativos e rendimentos dos fundos do mercado financeiro.
2. Desenvolva uma estimação por intervalo de confiança de 95% da média de ativos, do rendimento médio em sete dias e do rendimento médio em 30 dias para a população de fundos do mercado financeiro. Apresente uma interpretação administrativa de cada estimação por intervalo.
3. Discuta a implicação de suas conclusões em termos de como a BIS poderia utilizar esse tipo de informação ao preparar seu boletim semanal.
4. Quais outras informações você recomendaria à BIS coletar a fim de oferecer a informação mais útil aos seus clientes?

Estudo de Caso 2 – Gulf Real Estate Properties

A Gulf Real Estate Properties, Inc., é uma empresa imobiliária localizada no sudoeste da Flórida. A empresa, que divulga a si mesma como "especialista no mercado imobiliário", monitora as vendas em condomínios coletando dados sobre a localização, preço de tabela, preço de venda e número de dias necessários para vender cada unidade.

ARQUIVO DA INTERNET
Bock

Tabela 8.6 Dados da Bock Investiment Services

Fundo do Mercado Financeiro	Ativos (US$ milhões)	Rendimento (%) em 7 dias	Rendimento (%) em 30 dias
Amcore	103,9	4,10	4,08
Alger	156,7	4,79	4,73
Arch MM/Trust	496,5	4,17	4,13
BT Instit Treas	197,8	4,37	4,32
Benchmark Div	2.755,4	4,54	4,47
Bradford	707,6	3,88	3,83
Capital Cash	1,7	4,29	4,22
Cash Mgt Trust	2.707,8	4,14	4,04
Composite	122,8	4,03	3,91
Cowen Standby	694,7	4,25	4,19
Cortland	217,3	3,57	3,51
Declaration	38,4	2,67	2,61
Dreyfus	4.832,8	4,01	3,89
Elfun	81,7	4,51	4,41
FFB Cash	506,2	4,17	4,11
Federated Master	738,7	4,41	4,34

Tabela 8.6 Dados da Bock Investiment Services (continuação)

Fundo do Mercado Financeiro	Ativos (US$ milhões)	Rendimento (%) em 7 dias	Rendimento (%) em 30 dias
Fidelity Cash	13.272,8	4,51	4,42
Flex-fund	172,8	4,60	4,48
Fortis	105,6	3,87	3,85
Franklin Money	996,8	3,97	3,92
Freedom Cash	1.079,0	4,07	4,01
Galaxy Money	801,4	4,11	3,96
Government Cash	409,4	3,83	3,82
Hanover Cash	794,3	4,32	4,23
Heritage Cash	1.008,3	4,08	4,00
Infinity/Alpha	53,6	3,99	3,91
John Hancock	226,4	3,93	3,87
Landmark Funds	481,3	4,28	4,26
Liquid Cash	388,9	4,61	4,64
MarketWatch	10,6	4,13	4,05
Merrill Lynch Money	27.005,6	4,24	4,18
NCC Funds	113,4	4,22	4,20
Nationwide	517,3	4,22	4,14
Overland	291,5	4,26	4,17
Pierpont Money	1.991,7	4,50	4,40
Portico Money	161,6	4,28	4,20
Prudential MoneyMart	6.835,1	4,20	4,16
Reserve Primary	1408,8	3,91	3,86
Schwab Money	10.531,0	4,16	4,07
Smith Barney Cash	2.947,6	4,16	4,12
Stagecoach	1.502,2	4,18	4,13
Strong Money	470,2	4,37	4,29
Transamerica Cash	175,5	4,20	4,19
United Cash	323,7	3,96	3,89
Woodward Money	1.330,0	4,24	4,21

Fonte: Barron's, 3 de outubro de 1994.

Tabela 8.7 Dados de vendas da Gulf Real Estate Properties

Condomínios com Vista para o Golfo			Condomínios sem Vista para o Golfo		
Preço de Tabela	Preço de Venda	Dias Necessários para Vender	Preço de Tabela	Preço de Venda	Dias Necessários para Vender
495,0	475,0	130	217,0	217,0	182
379,0	350,0	71	148,0	135,5	338
529,0	519,0	85	186,5	179,0	122
552,5	534,5	95	239,0	230,0	150
334,9	334,9	119	279,0	267,5	169
550,0	505,0	92	215,0	214,0	58
169,9	165,0	197	279,0	259,0	110
210,0	210,0	56	179,9	176,5	130
975,0	945,0	73	149,9	144,9	149
314,0	314,0	126	235,0	230,0	114
315,0	305,0	88	199,8	192,0	120
885,0	800,0	282	210,0	195,0	61
975,0	975,0	100	226,0	212,0	146
469,0	445,0	56	149,9	146,5	137
329,0	305,0	49	160,0	160,0	281
365,0	330,0	48	322,0	292,5	63
332,0	312,0	88	187,5	179,0	48
520,0	495,0	161	247,0	227,0	52
425,0	405,0	149			

ARQUIVO DA INTERNET

Gulf Prop

Tabela 8.7 Dados de vendas da Gulf Real Estate Properties (continuação)

Condomínios com Vista para o Golfo		
Preço de Tabela	Preço de Venda	Dias Necessários para Vender
675,0	669,0	142
409,0	400,0	28
649,0	649,0	29
319,0	305,0	140
425,0	410,0	85
359,0	340,0	107
469,0	449,0	72
895,0	875,0	129
439,0	430,0	160
435,0	400,0	206
235,0	227,0	91
638,0	618,0	100
629,0	600,0	97
329,0	309,0	114
595,0	555,0	45
339,0	315,0	150
215,0	200,0	48
395,0	375,0	135
449,0	425,0	53
499,0	465,0	86
439,0	428,5	158

Cada condomínio é classificado como *Com Vista para o Golfo*, se estiver localizado diretamente defronte ao Golfo do México, ou *Sem Vista para o Golfo*, se estiver localizado na baía ou em um campo de golfe, próximo, mas não no Golfo. Dados amostrais do Multiple Listing Service de Naples, Flórida, forneceram dados de venda recentes de 40 condomínios *Com Vista para o Golfo* e 18 condomínios *Sem Vista para o Golfo*.[8] Os preços estão expressos em milhares de dólares. Os dados encontram-se na Tabela 8.7.

Relatório Administrativo

1. Use a estatística descritiva apropriada para sintetizar cada uma das três variáveis correspondentes aos 40 condomínios *Com Vista para o Golfo.*
2. Utilize a estatística descritiva apropriada para sintetizar cada uma das três variáveis correspondentes aos condomínios *Sem Vista para o Golfo.*
3. Compare os resultados de seu sumário estatístico. Discuta quaisquer resultados estatísticos específicos que possam ajudar um agente imobiliário a entender o mercado de condomínios.
4. Desenvolva uma estimação por intervalo de confiança de 95% da média populacional dos preços de venda e a média populacional do número de dias necessários para vender condomínios *Com Vista para o Golfo*. Interprete os resultados que obteve.
5. Estabeleça uma estimação por intervalo de confiança de 95% da média populacional dos preços de venda e a média populacional do número de dias necessários para vender condomínios *Sem Vista para o Golfo*. Interprete os resultados que obteve.
6. Suponha que o gerente de uma filial tenha solicitado estimativas do preço médio de venda de condomínios *Com Vista para o Golfo*, com uma margem de erro de US$ 40 mil, e o preço médio de venda de condomínios *Sem Vista para o Golfo*, com uma margem de erro de US$ 15 mil. Usando 95% de confiança, quais devem ser os tamanhos de amostra?
7. A Gulf Real Estate Properties assinou, há pouco tempo, contratos de duas novas intermediações de venda: um condomínio *Com Vista para o Golfo* com um preço de tabela de US$ 589 mil e um con-

[8] Dados baseados em vendas de condomínios publicados no *Multiple Listing Service* (MLS) de Naples (Coldwell Banker, junho de 2000).

domínio *Sem Vista para o Golfo* com um preço de tabela de US$ 285 mil. Qual é sua estimativa do preço de venda final e do número de dias necessários para vender cada uma dessas unidades?

Estudo de Caso 3 – Metropolitan Research, Inc.

A Metropolitan Research, Inc., uma organização de pesquisa de consumo, realiza pesquisas projetadas para avaliar ampla variedade de produtos e serviços disponíveis aos consumidores. Em um estudo em particular, a Metropolitan queria avaliar a satisfação do consumidor com o desempenho dos automóveis produzidos por uma grande montadora de Detroit. Um questionário enviado aos proprietários de carros completamente equipados produzidos pela montadora revelou diversas reclamações sobre problemas de transmissão prematuros. Para saber mais sobre as falhas de transmissão, a Metropolitan usou uma amostra de reparos de transmissão reais fornecida por uma firma de reparo de caixas de câmbio da região de Detroit. Os dados a seguir apresentam o número real de milhas rodadas de 50 veículos no momento em que ocorreu a falha de transmissão.

85.092	32.609	59.465	77.437	32.534	64.090	32.464	59.902
39.323	89.641	94.219	116.803	92.857	63.436	65.605	85.861
64.342	61.978	67.998	59.817	101.769	95.774	121.352	69.568
74.276	66.998	40.001	72.069	25.066	77.098	69.922	35.662
74.425	67.202	118.444	53.500	79.294	64.544	86.813	116.269
37.831	89.341	73.341	85.288	138.114	53.402	85.586	82.256
77.539	88.798						

ARQUIVO DA INTERNET
Auto

Relatório Administrativo

1. Use a estatística descritiva apropriada para resumir os dados de falha de transmissão.
2. Desenvolva um intervalo de confiança de 95% do número médio de milhas rodadas até o momento da falha de transmissão para a população de automóveis que apresentaram falhas de transmissão. Apresente uma interpretação gerencial da estimação por intervalo.
3. Discuta a implicação de sua conclusão estatística em termos da convicção de que alguns proprietários dos automóveis enfrentaram problemas de transmissão prematuros.
4. Quantos registros de reparos devem ser tomados como amostra se a empresa que realiza a pesquisa quiser que o número médio de milhas até a ocorrência da falha de transmissão seja estimado com uma margem de erro de 5 mil milhas? Use 95% de confiança.
5. Quais outras informações você gostaria de reunir para avaliar mais plenamente o problema de falhas de transmissão?

Apêndice 8.1 – Estimação por Intervalo com o Minitab

Descrevemos o uso do Minitab para construir intervalos de confiança de uma média populacional e de uma proporção populacional.

Média da População: σ Conhecido

Ilustramos a estimação por intervalo usando o exemplo da Lloyd's na Seção 8.1. As quantias gastas em cada ida às compras referentes à amostra de 100 clientes estão na coluna C1 de uma planilha do Minitab. Presume-se que o desvio padrão $\sigma = 20$ da população seja conhecido. As etapas a seguir podem ser usadas para calcular uma estimação por intervalo de confiança de 95% da média populacional.

ARQUIVO DA INTERNET
Lloyd's

Etapa 1. Selecione o menu **Stat**
Etapa 2. Escolha **Basic Statistics**
Etapa 3. Escolha **1-Sample Z**
Etapa 4. Quando a caixa de diálogo 1-Sample Z aparecer:
Digite C1 na caixa **Samples in columns**
Digite 20 na caixa **Standard deviation**
Etapa 5. Dê um clique em **OK**

O padrão do Minitab é um grau de confiança de 95%. Para especificar um grau de confiança diferente. por exemplo, 90%, acrescente o seguinte à etapa 4.

Selecione **Options**
Quando a caixa de diálogo 1-Sample Z-Options aparecer:
  Digite 90 na caixa **Confidence level**
Dê um clique em **OK**

Média da População: σ Desconhecido

ARQUIVO DA INTERNET

Ilustramos a estimação por intervalo usando os dados da Tabela 8.3, a qual exibe os saldos de cartões de crédito de uma amostra de 85 famílias. Os dados encontram-se na coluna C1 de uma planilha do Minitab. Nesse caso, a estimativa do desvio padrão σ da população será feita por meio do desvio padrão σ da amostra. As etapas a seguir podem ser usadas para calcular uma estimação por intervalo de confiança de 95% da média populacional.

Etapa 1. Selecione o menu **Stat**
Etapa 2. Escolha **Basic Statistics**
Etapa 3. Escolha **1-Sample t**
Etapa 4. Quando a caixa de diálogo 1-Sample t aparecer:
  Digite C1 na caixa **Samples in columns**
Etapa 5. Dê um clique em **OK**

O padrão do Minitab é um grau de confiança de 95%. Para especificar um grau de confiança diferente, por exemplo, 90%, acrescente o seguinte à etapa 4.

Selecione **Options**
Quando a caixa de diálogo 1-Sample t-Options aparecer:
  Digite 90 na caixa **Confidence level**
Dê um clique em **OK**

Proporção da População

ARQUIVO DA INTERNET

Ilustramos a estimação por intervalo usando os dados de pesquisa de mulheres golfistas apresentados na Seção 8.4. Os dados encontram-se na coluna C1 de uma planilha do Minitab. As respostas estão registradas como Sim se a golfista estiver satisfeita com a disponibilidade de *tee times* e Não se não estiver. As etapas a seguir podem ser usadas para calcular uma estimação por intervalo de confiança de 95% da proporção de mulheres golfistas que estão satisfeitas com a disponibilidade de *tee times*.

Etapa 1. Selecione o menu **Stat**
Etapa 2. Escolha **Basic Statistics**
Etapa 3. Escolha **1 Proportion**
Etapa 4. Quando a caixa de diálogo **1 Proportion** aparecer:
  Digite C1 na caixa **Samples in columns**
Etapa 5. Selecione **Options**
Etapa 6. Quando a caixa de diálogo 1 Proportion-Options aparecer:
  Selecione **Use test and interval based on normal distribution**
  Dê um clique em **OK**
Etapa 7. Dê um clique em **OK**

O padrão do Minitab é um grau de confiança de 95%. Para especificar um grau de confiança diferente, por exemplo, 90%, digite 90 na caixa **Confidence Level**, quando a caixa de diálogo 1 Proportion-Options aparecer na etapa 6.

Nota: A rotina 1 Proportion do Minitab usa uma classificação em ordem alfabética das respostas e seleciona a *segunda resposta* para a proporção populacional de interesse. No exemplo das mulheres golfistas, o Minitab usou uma classificação em ordem alfabética Não-Sim e depois forneceu o intervalo de confiança relativo às respostas Sim. Uma vez que Sim era a resposta de interesse, a saída de dados (*output*) do Minitab foi ótima. Entretanto, se a classificação em ordem alfabética do Minitab não produzir a resposta de interesse, selecione qualquer célula da coluna e use a seqüência: Editor > Column > Value Order. Isso

lhe apresentará a opção de digitar uma ordem especificada pelo usuário, mas você deve listar a resposta de interesse em segundo lugar na caixa "define-an-order".

Apêndice 8.2 – Estimação por Intervalo com o Excel

Descrevemos o uso do Excel para construir intervalos de confiança de uma média populacional e de uma proporção populacional.

ARQUIVO DA INTERNET
Lloyds

Média da População: σ Conhecido

Ilustramos a estimação por intervalo usando o exemplo da Lloyd's na Seção 8.1. Presume-se que o desvio padrão σ =20 da população seja conhecido. As quantias gastas pela amostra de 100 clientes encontram-se na coluna A de uma planilha do Excel. As etapas a seguir podem ser usadas para calcular a margem de erro de uma estimativa da média populacional. Iniciamos utilizando a ferramenta Estatística Descritiva do Excel, apresentada no Capítulo 3.

Etapa 1. Selecione o menu **Ferramentas**
Etapa 2. Escolha **Análise de Dados**
Etapa 3. Escolha **Estatística Descritiva** na lista Ferramentas de Análise
Etapa 4. Quando a caixa Estatística Descritiva aparecer:
Digite A1:A101 na caixa **Intervalo de Entrada**
Selecione **Agrupado por Colunas**
Selecione **Rótulos na Primeira Linha**
Selecione **Intervalo de Saída**
Digite C1 na caixa **Intervalo de Saída**
Selecione **Resumo Estatístico**
Dê um clique em **OK**

O resumo estatístico aparecerá nas colunas C e D. Prossiga, calculando a margem de erro com o uso da função INT.CONFIANÇA do Excel da seguinte maneira:

Etapa 5. Selecione a célula C16 e digite o rótulo Margem de Erro
Etapa 6. Selecione a célula D16 e digite a fórmula =INT.CONFIANÇA(0,05;20;100) do Excel

Os três argumentos da função INT.CONFIANÇA são:
Alfa = 1 – coeficiente de confiança = 1 – 0,95 = 0,05
O desvio padrão da população = 20
O tamanho da amostra = 100 (*Nota*: Esse argumento aparece como COUNT.NÚM na célula D15.)

A estimação por ponto da média populacional encontra-se na célula D3, e a margem de erro encontra-se na célula D16. A estimação por ponto (82) e a margem de erro (3,92) permitem que o intervalo de confiança relativo à média populacional seja facilmente calculado.

ARQUIVO DA INTERNET
Balance

Média da População: σ Desconhecido

Ilustramos a estimação por intervalo usando os dados da Tabela 8.3, a qual apresenta os saldos de cartões de crédito de uma amostra de 85 famílias. Os dados encontram-se na coluna A de uma planilha do Excel. As etapas apresentadas a seguir podem ser usadas para calcular a estimação por ponto e a margem de erro da estimação por intervalo de uma média populacional. Usaremos a ferramenta Estatística Descritiva apresentada no Capítulo 3.

Etapa 1. Selecione o menu **Ferramentas**
Etapa 2. Escolha **Análise de Dados**
Etapa 3. Escolha **Estatística Descritiva** na lista Ferramentas de Análise
Etapa 4. Quando a caixa Estatística Descritiva aparecer:
Digite A1:A86 na caixa **Intervalo de Entrada**
Selecione **Agrupado por Colunas**
Selecione **Rótulos na Primeira Linha**

Selecione **Intervalo de Saída**
Digite C1 na caixa Intervalo de Saída
Selecione **Resumo Estatístico**
Selecione **Nível de Confiabilidade para Média**
Digite 95 na caixa Nível de Confiabilidade para Média
Dê um clique em **OK**

O resumo estatístico aparecerá nas colunas C e D. A estimação por ponto da média populacional aparece na célula D3. A margem de erro, rotulada como "Nível de Confiabilidade(95,0%)", aparecerá na célula D16. A estimação por ponto (US$ 5.900) e a margem de erro (US$ 660) permitem que o intervalo de confiança relativo à média populacional seja facilmente calculado. Os dados de saída (*output*) desse procedimento do Excel encontram-se na Figura 8.10.

Proporção da População

Ilustramos a estimação por intervalo usando os dados de pesquisa de mulheres golfistas apresentados na Seção 8.4. Os dados encontram-se na coluna A de uma planilha do Excel. As respostas individuais estão registradas como Sim se a golfista estiver satisfeita com a disponibilidade de *tee times*, e como Não se não estiver. O Excel não oferece uma rotina própria para manipular a estimação de uma proporção populacional; entretanto, é relativamente fácil de desenvolver um modelo (*template*) Excel que possa ser usado para essa finalidade. O modelo mostrado na Figura 8.11 fornece uma estimação por intervalo de confiança de 95% da proporção de mulheres golfistas que estão satisfeitas com a disponibilidade de *tee times*. Observe que a planilha em segundo plano na Figura 8.11 exibe fórmulas nas células que produzem os resultados de estimação por intervalo apresentados na planilha que está em primeiro plano. As etapas a seguir são necessárias para que se possa usar o modelo para esse conjunto de dados (*data set*).

Figura 8.10 Estimação por intervalo da média da população de cartões de crédito com o Excel

Nota: As linhas do intervalo 18 a 80 estão ocultas.

	A	B	C	D	E	F
1	Saldo		*Saldo*			
2	9619					
3	5364		Média	5900	Estimação por Ponto	
4	8348		Erro Padrão	331.7		
5	7348		Mediana	5759		
6	381		Moda	8047		
7	2998		Desvio Padrão	3058		
8	1686		Variância da Amostra	9351364		
9	1962		Curtose	0.2327		
10	4920		Assimetria	0.4076		
11	5047		Intervalo	14061		
12	6921		Mínimo	381		
13	5759		Máximo	14442		
14	8047		Soma	501500		
15	3924		Contagem	85		
16	3470		Grau de Confiança (95,0%)	660	Margem de Erro	
17	5994					
81	5938					
82	5266					
83	10658					
84	3910					
85	7503					
86	1582					
87						

ARQUIVO DA INTERNET
Interval p

Etapa 1. Digite o intervalo de dados A2:A901 na fórmula =CONT.VALORES da célula D3
Etapa 2. Digite Sim como a resposta de interesse na célula D4
Etapa 3. Digite o intervalo de dados A2:A901 na fórmula =CONT.SE da célula D5
Etapa 4. Digite 0,95 como coeficiente de confiança na célula D8

O modelo fornece automaticamente o intervalo de confiança nas células D15 e D16.

Esse modelo pode ser usado para calcular o intervalo de confiança relativo a uma proporção populacional de outras aplicações. Por exemplo, para calcular a estimação por intervalo de um novo conjunto de dados, digite os novos dados amostrais na coluna A da planilha e depois faça as alterações nas quatro células, conforme mostrado. Se os novos dados amostrais já tiverem sido resumidos, eles não precisam ser introduzidos na planilha. Nesse caso, digite o tamanho da amostra na célula D3 e a proporção da amostra na célula D6; o modelo de planilha produzirá então o intervalo de confiança da proporção populacional. A planilha da Figura 8.11 está disponível no arquivo Interval p no CD anexo a este livro.

Figura 8.11 Modelo Excel de estimação por intervalo de uma proporção populacional

	A	B	C	D	E
1	Resposta		Estimação por Intervalo de uma Proporção Populacional		
2	Sim				
3	Não		Tamanho da Amostra	=CONT.VALORES(A2:A901)	
4	Sim		Resposta de Interesse	Sim	
5	Sim		Contagem e Respostas	=CONT.SE(A2:A901;D4)	
6	Não		Proporção da Amostra	=D5/D3	
7	Não				
8	Não		Coeficiente de confiança	0.95	
9	Sim		Valor z	=INV.NORM(0,51D8/2)	
10	Sim				
11	Sim		Desvio Padrão	=RAIZ(D6*(1-D6/D3)	
12	Não		Margem de Erro	=D9*D11	
13	Não				
14	Sim		Estimação por Ponto	=D6	
15	Não		Limite Mínimo	=D14-D12	
16	Não		Limite Máximo	=D14+D12	
17	Sim				
18	Não				
901	Sim				
902					

	A	B	C	D	E	F	G
1	Resposta		Estimação por Intervalo de uma Proporção Populacional				
2	Sim						
3	Não		Tamanho da Amostra	900			
4	Sim		Resposta de Interesse	Sim	Digite a resposta de interesse		
5	Sim		Contagem e Respostas	396			
6	Não		Proporção da Amostra	0,4400			
7	Não						
8	Não		Coeficiente de confiança	0,95	Digite o coeficiente de confiança		
9	Sim		Valor z	1,960			
10	Sim						
11	Sim		Desvio Padrão	0,0165			
12	Não		Margem de Erro	0,0324			
13	Não						
14	Sim		Estimação por Ponto	0,4400			
15	Não		Limite Mínimo	0,4076			
16	Não		Limite Máximo	0,4724			
17	Sim						
18	Não						
901	Sim						
902							

Nota: As linhas do intervalo 19 a 900 estão ocultas.

CAPÍTULO 9

Testes de Hipóteses

ESTATÍSTICA NA PRÁTICA

JOHN MORRELL & COMPANY*
Cincinnati, Ohio

A John Morrell & Company, que se iniciou na Inglaterra em 1827, é considerada o mais antigo frigorífico em operação contínua nos Estados Unidos. Ela é uma subsidiária integral da Smithfield Foods, de Smithfield, Virgínia, gerenciada independentemente. A John Morrell & Company oferece uma extensa linha de produtos frigoríficos e de carne de porco fresca a consumidores de 13 marcas regionais, entre as quais se incluem a John Morrell, E-Z-Cut, Tobin's First Prize, Dinner Bell, Hunter, Kretschmar, Rath, Rodeo, Shenson, Farmers Hickory Brand, Iowa Quality e Peyton's. Cada marca regional desfruta de elevado reconhecimento de marca e fidelidade entre os seus consumidores.

Pesquisas de mercado realizadas pela Morrell fornecem à administração informações atualizadas sobre os vários produtos da empresa e como estes se comparam com as marcas concorrentes de produtos similares. Um estudo recente investigou a preferência pela linha Convenient Cuisine Beef Pot Roast, da Morrell, em comparação com produtos frigoríficos de duas grandes concorrentes. No teste comparativo de três produtos, foi utilizada uma amostra de consumidores que deveriam indicar qual era a avaliação dos produtos em termos de sabor, aspecto, aroma e preferência global.

Uma das preocupações da pesquisa era se a linha Convenient Cuisine Beef Pot Roast, da Morrell, era a opção preferida de mais de 50% da população consumidora. Admitindo-se que p indique a proporção da população que preferia o produto da Morrell, o teste de hipóteses referente à questão da pesquisa é:

$$H_0: p \leq 0{,}50$$
$$H_a: p > 0{,}50$$

* Os autores agradecem a Marty Butler, vice-presidente de Marketing da John Morrell & Company, por fornecer esta "Estatística na Prática".

A hipótese nula H_0 indica que a preferência pelo produto da Morrell é menor ou igual a 50%. Se os dados amostrais sustentarem a opção de rejeitar H_0 em favor da hipótese alternativa H_a, a Morrell chegará à conclusão com base em pesquisa de que, em uma comparação de três produtos, seus produtos são os preferidos por mais de 50% da população consumidora.

Em um estudo independente de testes de sabor envolvendo uma amostra de 224 consumidores de Cincinnati, Milwaukee e Los Angeles, 150 consumidores escolheram o Convenient Cuisine Beef Pot Roast, da Morrell, como o produto preferido. Com o uso de procedimentos estatísticos de teste de hipóteses, a hipótese nula H_0 foi rejeitada. O estudo produziu evidências estatísticas que dão suporte a H_a e à conclusão segundo a qual o produto da Morrell é o preferido por mais de 50% da população consumidora.

A estimação por ponto da proporção populacional foi $\bar{p} = 150/224 = 0,67$. Desse modo, os dados da amostra forneceram o suporte para que a empresa promovesse um anúncio de revista mostrando que, em uma comparação do sabor de três produtos, o Convenient Cuisine Beef Pot Roast, da Morrell, "obtinha uma preferência de 2 por 1 sobre a concorrência".

Neste capítulo, discutiremos como formular hipóteses e como realizar testes idênticos ao que é utilizado pela Morrell. Por meio de análise dos dados amostrais seremos capazes de determinar se a hipótese deve ou não ser rejeitada.

Nos Capítulos 7 e 8, informamos como uma amostra pode ser usada para desenvolver estimações por ponto e por intervalo dos parâmetros populacionais. Neste capítulo, prosseguiremos a discussão da inferência estatística, mostrando como o teste de hipóteses pode ser usado para determinar se uma afirmação sobre o valor de um parâmetro populacional deve ou não ser rejeitada.

Ao testar hipóteses, iniciamos por criar uma hipótese experimental a respeito de um parâmetro da população. Essa hipótese experimental é chamada **hipótese nula**. Ela é denotada por H_0. Definimos, então, outra hipótese, denominada **hipótese alternativa**, a qual é o oposto daquilo que é formulado na hipótese nula. A hipótese alternativa é denotada por H_a. O procedimento de teste de hipóteses usa dados de uma amostra para testar as duas afirmações antagônicas, indicadas por H_0 e H_a.

Este capítulo mostra como se pode realizar testes de hipóteses a respeito de uma média populacional e de uma proporção populacional. Iniciaremos apresentando exemplos que ilustram critérios para o desenvolvimento das hipóteses nula e alternativa.

9.1 COMO DESENVOLVER AS HIPÓTESES NULA E ALTERNATIVA

Aprender a formular hipóteses corretamente é algo que demandará prática. Aguarde certa confusão inicial a respeito da escolha adequada de H_0 e de H_a. Os exemplos desta seção apresentam uma série de formas de H_0 e H_a, dependendo da aplicação.

Em algumas aplicações pode não ser claro à primeira vista como as hipóteses nula e alternativa devem ser formuladas. Deve-se tomar cuidado para estruturar as hipóteses de maneira apropriada a fim de que a conclusão do teste de hipóteses produza a informação que o pesquisador ou o tomador de decisão desejam. Diretrizes para estabelecer as hipóteses nula e alternativa serão dadas para três tipos de situação nas quais comumente se empregam procedimentos de teste de hipóteses.

Como Testar Hipóteses de Pesquisa

Considere um modelo de automóvel em particular que atinge atualmente a eficiência média de 10,21 km/L em termos de consumo de combustível. Uma equipe de pesquisa de produto desenvolveu um novo sistema de injeção de combustível projetado especificamente para aumentar a taxa de quilômetros por litro. Para avaliar o novo sistema, diversas unidades serão produzidas, instaladas em automóveis e submetidas a testes de direção controlados. Aqui, a equipe de pesquisa de produto está à procura de comprovação de que o novo sistema *aumenta* a taxa média de quilômetros por litro. Nesse caso, a hipótese de pesquisa é que o novo sistema de injeção de combustível produzirá uma taxa média de quilômetros por litro superior a 10,21; ou seja, $\mu > 10,21$. Como diretriz geral, uma hipótese de pesquisa deve ser formulada como a *hipótese alternativa*. Portanto, as hipóteses nula e alternativa relativas ao estudo são:

$$H_0: \mu \leq 10,21$$
$$H_a: \mu > 10,21$$

Se os resultados da amostra indicarem que H_0 não pode ser rejeitada, os pesquisadores não poderão concluir que o novo sistema de injeção de combustível é melhor. Talvez, mais pesquisas e testes subseqüentes devam ser realizados. Entretanto, se os resultados da amostra indicarem que H_0 pode ser rejeitada, os pesquisadores poderão inferir que H_a: $\mu > 10{,}21$ é verdadeira. Com essa conclusão, os pesquisadores obtêm a base estatística necessária para afirmar que o novo sistema aumenta o número médio de quilômetros por litro. Portanto, a produção com o novo sistema deve ser considerada.

A conclusão de que a hipótese de pesquisa pode ser verdadeira é obtida se os dados da amostra contradisserem a hipótese nula.

Em estudos de pesquisa desse tipo, as hipóteses nula e alternativa devem ser formuladas de tal maneira que a rejeição de H_0 corrobore a conclusão da pesquisa. As hipóteses de pesquisa, portanto, devem ser expressas como a hipótese alternativa.

Como Testar a Validade de uma Afirmação

Como uma ilustração do teste de validade de uma afirmação, considere a situação em que um fabricante de refrigerantes declara que os frascos de dois litros dos seus produtos contêm, no mínimo, uma média de 1,99 L. Uma amostra de frascos de dois litros será selecionada e o conteúdo, medido, para testar a afirmação do fabricante. Nesse tipo de teste de hipóteses, geralmente presumimos que a afirmação do fabricante é verdadeira, a menos que a evidência da amostra seja contraditória. Usando esse critério no exemplo dos frascos de refrigerante, afirmaríamos que as hipóteses nula e alternativa são as seguintes:

$$H_0: \mu \geq 1{,}99$$
$$H_a: \mu < 1{,}99$$

Se os resultados da amostra indicarem que H_0 não pode ser rejeitada, a afirmação do fabricante não será contestada. Entretanto, se os resultados da amostra indicarem que H_0 pode ser rejeitada, a inferência é que H_a: $\mu < 1{,}99$ é verdadeira. Com essa conclusão, a evidência estatística indica que a afirmação do fabricante é incorreta e que os frascos de refrigerante são preenchidos com uma média menor que a anunciada quantidade de 1,99 L. As medidas cabíveis contra o fabricante devem ser consideradas.

Geralmente se concede o benefício da dúvida à afirmação do fabricante, e ela é estabelecida como a hipótese nula. A conclusão de que a afirmação é falsa pode ser feita se a hipótese nula for rejeitada.

Em situações que envolvem testar a validade de uma afirmação, a hipótese nula geralmente se baseia no pressuposto de que a afirmação é verdadeira. A hipótese alternativa é então formulada a fim de que a rejeição de H_0 produza a evidência estatística de que a hipótese declarada é incorreta. Iniciativas para corrigir a afirmação devem ser consideradas sempre que H_0 for rejeitada.

Como Testar em Situações de Tomada de Decisão

Quando se testam hipóteses de pesquisa ou a validade de uma afirmação, as providências necessárias são postas em prática se H_0 for rejeitada. Em muitos casos, no entanto, devem-se tomar providências tanto quando H_0 não pode ser rejeitada como quando H_0 pode ser rejeitada. Em geral, esse tipo de situação ocorre quando um tomador de decisão precisa escolher entre dois cursos de ação: um associado à hipótese nula e outro, à hipótese alternativa. Por exemplo, considerando uma amostra de peças de uma remessa recém-recebida, um inspetor de controle da qualidade precisa decidir se aceitará a remessa ou se a devolverá ao fornecedor porque ela não cumpre as especificações. Suponha que as especificações de uma peça em particular exijam um tamanho médio de duas polegadas por peça. Se o tamanho médio for maior ou menor que o padrão de duas polegadas, as peças causarão problemas de qualidade na operação de montagem. Nesse caso, as hipóteses nula e alternativa serão formuladas da seguinte maneira:

$$H_0: \mu = 2$$
$$H_a: \mu \neq 2$$

Se os resultados da amostra indicarem que H_0 não pode ser rejeitada, o inspetor de controle da qualidade não terá nenhuma razão para duvidar de que a remessa esteja de acordo com as especificações, e a remessa será aceita. Entretanto, se os resultados da amostra indicarem que H_0 deve ser rejeitada, a conclusão será de que as peças não cumprem as especificações. Nesse caso, o inspetor de controle da qualidade terá suficientes evidências para devolver a remessa ao fornecedor. Desse modo, vemos que para esses tipos de situação, providências devem ser tomadas tanto quando H_0 não pode ser rejeitada como quando H_0 pode ser rejeitada.

Resumo das Formas das Hipóteses Nula e Alternativa

Os testes de hipóteses deste capítulo envolvem dois parâmetros populacionais: a média populacional e a proporção populacional. Dependendo da situação, o teste de hipóteses a respeito de um parâmetro popu-

lacional pode assumir uma das três formas possíveis: duas delas usam desigualdades na hipótese nula; a terceira utiliza uma igualdade na hipótese nula. Em relação aos testes de hipóteses, que envolvem uma média populacional, admitimos que μ_0 denota o valor hipotético, e então precisamos escolher uma das três formas seguintes para o teste da hipótese:

As três formas possíveis das hipóteses H_0 e H_a são apresentadas aqui. Observe que o sinal de igualdade sempre aparece na hipótese nula H_0.

$$H_0: \mu \geq \mu 0 \qquad H_0: \mu \leq \mu 0 \qquad H_0: \mu = \mu 0$$
$$H_a: \mu < \mu 0 \qquad H_a: \mu > \mu 0 \qquad H_a: \mu \neq \mu 0$$

Por razões que se tornarão claras adiante, as duas primeiras formas são chamadas testes unicaudais. A terceira forma é denominada teste bicaudal.

Em muitas situações, a escolha de H_0 e H_a não é clara, e é necessário discernimento para selecionar a forma apropriada. Entretanto, como mostram as formas apresentadas anteriormente, o termo de igualdade ($\geq$, $\leq$ ou $=$) *sempre* aparece na hipótese nula.

Ao selecionar a forma apropriada de H_0 e H_a, tenha em mente que a hipótese alternativa freqüentemente é aquilo que o teste tenta estabelecer. Portanto, perguntar se o usuário está à procura de evidências que apóiem a $\mu < \mu_0$, $\mu > \mu_0$ ou $\mu \neq \mu_0$ ajudará a determinar H_a. Os exercícios a seguir foram idealizados para que você adquira prática na escolha da forma apropriada do teste de hipóteses envolvendo uma média populacional.

Exercícios

1. O gerente do Danvers-Hilton Resort Hotel afirmou que o valor médio da conta dos hóspedes em um final de semana é igual a US$ 600 ou menos. Um membro da equipe de contabilidade do hotel observou que o total cobrado nas contas dos hóspedes se elevou nos últimos meses. O contador usará uma amostra de contas de hóspedes em fins de semana para testar a afirmação do gerente.
   a. Qual forma de hipótese deve ser usada para testar a afirmação do gerente? Explique.
   $$H_0: \mu \geq 600 \qquad H_0: \mu \leq 600 \qquad H_0: \mu = 600$$
   $$H_a: \mu < 600 \qquad H_a: \mu > 600 \qquad H_a: \mu \neq 600$$
   b. Qual conclusão é apropriada quando H_0 não pode ser rejeitada?
   c. Qual conclusão é apropriada quando H_0 pode ser rejeitada?

AUTOTESTE

2. O gerente de uma concessionária de automóveis está pensando em um novo plano de bonificações para aumentar o volume de vendas. Atualmente, o volume médio de vendas é de 14 automóveis por mês. O gerente quer realizar um estudo e pesquisa para verificar se o novo plano de bonificações aumenta o volume de vendas. Para coletar dados sobre o plano, uma amostra da equipe de vendas será autorizada a vender sob o novo plano de bonificação durante o período de um mês.
   a. Desenvolva as hipóteses nula e alternativa mais apropriadas a essa situação de pesquisa.
   b. Comente a conclusão relativa a quando H_0 não pode ser rejeitada.
   c. Comente a conclusão relativa a quando H_0 pode ser rejeitada.
3. Uma operação de linha de produção foi projetada para encher caixas de sabão em pó com um peso médio de 0,907 kg. Uma amostra das caixas é selecionada periodicamente e pesada para determinar se há a ocorrência de enchimentos abaixo ou acima do padrão. Se os dados da amostra levarem à conclusão de que há enchimentos abaixo ou acima do padrão, a linha de produção será interrompida e ajustada para se obter o enchimento apropriado.
   a. Formule as hipóteses nula e alternativa que ajudem a decidir se a linha de produção deve ser interrompida e ajustada.
   b. Comente a conclusão e a decisão de quando H_0 não pode ser rejeitada.
   c. Comente a conclusão e a decisão de quando H_0 pode ser rejeitada.
4. Em virtude do tempo e dos custos elevados de produção e transformação, um diretor de manufatura precisa convencer a administração de que um novo método de manufatura proposto reduz os custos, antes de o novo método ser implementado. O método de produção atual opera com um custo médio de US$ 220 por hora. Um estudo e pesquisa medirão o custo do novo método ao longo de um período de produção amostral.
   a. Desenvolva as hipóteses alternativa e nula mais apropriadas a esse estudo.
   b. Comente a conclusão de quando H_0 não pode ser rejeitada.
   c. Comente a conclusão de quando H_0 pode ser rejeitada.

9.2 ERROS DO TIPO I E DO TIPO II

As hipóteses nula e alternativa são afirmações excludentes a respeito da população. Ou a hipótese nula H_0 é verdadeira ou a hipótese alternativa H_a é verdadeira, mas não ambas. Idealmente, o procedimento de teste de hipóteses deve levar à aceitação de H_0 quando H_0 é verdadeira, e à rejeição de H_0 quando H_a é verdadeira.

Tabela 9.1 Erros e conclusões corretas no teste de hipóteses

		Situação da População	
		H_0 verdadeira	H_a verdadeira
Conclusão	Aceitar H_0	Conclusão Correta	Erro do Tipo II
	Rejeitar H_0	Erro do Tipo I	Conclusão Correta

Infelizmente, as conclusões corretas nem sempre são possíveis. Uma vez que os testes de hipótese baseiam-se em informações de amostras, devemos admitir a possibilidade de erros. A Tabela 9.1 ilustra os dois tipos de erro que podem ser cometidos no teste de hipóteses.

A primeira linha da Tabela 9.1 revela o que pode acontecer se a conclusão for aceitar H_0. Se H_0 for verdadeira, essa conclusão está correta. Entretanto, se H_a for verdadeira, cometemos um **erro do Tipo II**; ou seja, aceitamos H_0 quando ela é falsa. A segunda linha mostra o que pode acontecer se a conclusão for rejeitar H_0. Se H_0 for verdadeira, cometemos um **erro do Tipo I**; ou seja, rejeitamos H_0 quando ela é verdadeira. Entretanto, se H_a for verdadeira, rejeitar H_0 será a ação correta.

Lembre-se da ilustração do teste de hipótese discutida na Seção 9.1, na qual uma equipe de pesquisa de produtos automobilísticos desenvolveu um novo sistema de injeção de combustível projetado para aumentar a taxa de quilômetros por litro de um automóvel em particular. Com o modelo atual que obtém uma média de 10,21 quilômetros por litro, a hipótese foi formulada da seguinte maneira:

$$H_0: \mu \leq 10{,}21$$
$$H_a: \mu > 10{,}21$$

A hipótese alternativa, H_a: $\mu > 10{,}21$, indica que os pesquisadores estão à procura de evidências amostrais que sustentem a conclusão de que a média populacional de quilômetros por litro com o novo sistema de injeção de combustível é superior a 10,21.

Nessa aplicação, o erro do Tipo I de rejeitar H_0 quando ela é verdadeira corresponde aos pesquisadores afirmarem que o novo sistema melhora a taxa de quilômetros por litro ($\mu > 10{,}21$) quando, de fato, o novo sistema não é melhor que o sistema atual. Em contrapartida, o erro do Tipo II de aceitar H_0 quando ela é falsa corresponde aos pesquisadores concluírem que o novo sistema não é melhor que o sistema atual ($\mu \leq 10{,}21$) quando, de fato, o novo sistema melhora o desempenho de quilômetros por litro.

Em relação ao teste da taxa de quilômetros por litro, a hipótese nula é H_0: $\mu \leq 10{,}21$. Suponha que a hipótese nula seja verdadeira enquanto igualdade; ou seja, $\mu = 10{,}21$. A probabilidade de cometer um erro do Tipo I quando a hipótese nula é verdadeira é chamada **nível de significância.** Desse modo, em relação ao teste de hipóteses da taxa de quilômetros por litro, o nível de significância é a probabilidade de se rejeitar H_0: $\mu \leq 10{,}21$ quando $\mu = 10{,}21$. Por causa da importância desse conceito, reformulamos agora a definição de nível de significância.

NÍVEL DE SIGNIFICÂNCIA

O nível de significância é a probabilidade de cometermos um erro do Tipo I quando a hipótese nula é verdadeira enquanto igualdade.

O símbolo grego α (alfa) é usado para denotar o nível de significância, e as escolhas habituais para α são 0,05 e 0,01.

Na prática, a pessoa que realiza o teste de hipóteses especifica o nível de significância. Ao selecionar α, essa pessoa controla a probabilidade de cometer um erro do Tipo I. Se o custo de cometer um erro do Tipo I for alto, valores pequenos de α são preferíveis. Se o custo de cometer um erro do Tipo I não for

Se os dados amostrais forem coerentes com a hipótese nula H_0, seguiremos a prática de optar pela conclusão "não rejeitar H_0." Essa conclusão é preferível a "aceitar H_0", porque a conclusão de aceitar H_0 nos coloca em risco de cometer um erro do Tipo II.

alto, valores maiores de α tipicamente são usados. Aplicações de testes de hipótese que somente controlam o erro do Tipo I freqüentemente são chamadas *testes de significância*. A maioria das aplicações de testes de hipótese é desse tipo.

Não obstante as aplicações de testes de hipóteses controlem a probabilidade de cometer um erro do Tipo I, elas nem sempre controlam a probabilidade de se cometer um erro do Tipo II. Portanto, se decidimos aceitar H_0 não poderemos determinar quão confiantes podemos estar a respeito dessa decisão. Em razão da incerteza associada à probabilidade de cometer um erro do Tipo II quando se realizam testes de significância, os estatísticos freqüentemente recomendam que devemos usar a afirmação "não rejeitar H_0" em vez de "aceitar H_0". O uso da afirmação "não rejeitar H_0" transmite a recomendação de se manter tanto o julgamento como a ação. Com efeito, ao não aceitar diretamente H_0, o estatístico evita o risco de cometer um erro do Tipo II. Quando quisermos que a probabilidade de cometer um erro do Tipo II não seja determinada e controlada, não faremos a afirmação "aceitar H_0". Nesses casos, somente duas conclusões são possíveis: *não rejeitar* H_0 ou *rejeitar* H_0. Embora o controle de um erro do Tipo II em testes de hipóteses não seja comum, ele pode ser feito. Livros mais avançados descrevem procedimentos para determinar e controlar a probabilidade de cometer um erro do Tipo II.* Se os controles apropriados tiverem sido estabelecidos para esse tipo de erro, ações baseadas na conclusão "aceitar H_0" podem ser apropriadas.

Exercícios

AUTOTESTE

5. A Nielsen divulgou que os jovens dos Estados Unidos assistem a 56,2 minutos de TV diariamente no horário nobre (*The Wall Street Journal Europe*, 18 de novembro de 2003). Um pesquisador acredita que os jovens alemães do sexo masculino passam mais tempo assistindo à TV no horário nobre. Uma amostra de jovens da Alemanha será selecionada pelo pesquisador, e o tempo que eles passam assistindo à TV em um dia será registrado. Os resultados da amostra serão usados para testar as hipóteses nula e alternativa seguintes:

$$H_0: \mu \leq 56{,}2$$
$$H_a: \mu > 56{,}2$$

   a. Qual é o erro de Tipo I nessa situação? Quais são as conseqüências de cometer esse erro?
   b. Qual é o erro de Tipo II nessa situação? Quais são as conseqüências de cometer esse erro?

6. O rótulo de um frasco de 2,83 litros de suco de laranja afirma que o suco de laranja contém em média 1 grama ou menos de gordura. Responda às questões a seguir considerando um teste de hipóteses que possa ser usado para testar a afirmação constante no rótulo.
   a. Desenvolva as hipóteses nula e alternativa apropriadas.
   b. Qual é o erro de Tipo I nessa situação? Quais são as conseqüências de cometer esse erro?
   c. Qual é o erro de Tipo II nessa situação? Quais são as conseqüências de cometer esse erro?

7. A equipe de vendas da Carpetland atinge uma média de US$ 8 mil em vendas por semana. Steve Contois, o vice-presidente da firma, propôs um programa de remuneração com novos incentivos de vendas. Steve espera que os resultados de um período experimental de vendas lhe possibilitem concluir que o programa de remuneração aumenta a média de vendas por vendedor.
   a. Desenvolva as hipóteses nula e alternativa apropriadas.
   b. Qual é o erro de Tipo I nessa situação? Quais são as conseqüências de cometer esse erro?
   c. Qual é o erro de Tipo II nessa situação? Quais são as conseqüências de cometer esse erro?

8. Suponha que um novo método de produção seja implementado se um teste de hipóteses sustentar a conclusão de que o novo método reduz a média de custo operacional por hora.
   a. Estabeleça as hipóteses nula e alternativa apropriadas considerando que o custo médio do método de produção atual seja igual a US$ 220 por hora.
   b. Qual é o erro de Tipo I nessa situação? Quais são as conseqüências de cometer esse erro?
   c. Qual é o erro de Tipo II nessa situação? Quais são as conseqüências de cometer esse erro?

*Veja, por exemplo, *Statistics for Business and Economics*, 9. ed., de ANDERSON, D. R. et al. (Cincinnati: South-Western, 2005).

9.3 MÉDIA DA POPULAÇÃO: σ CONHECIDO

No Capítulo 8, dissemos que o caso em que σ é conhecido corresponde a aplicações nas quais dados históricos ou outras informações estão disponíveis e que nos possibilitam obter uma boa estimativa do desvio padrão da população antes da amostragem. Nesses casos, o desvio padrão da população pode, para todos os efeitos, ser considerado conhecido. Nesta seção, mostramos como realizar um teste de hipóteses sobre a média populacional, considerando o caso em que σ seja conhecido.

Os métodos apresentados nesta seção são exatos se a amostra for selecionada de uma população que está normalmente distribuída. Nos casos em que não é razoável supormos que a população esteja normalmente distribuída, ainda assim esses métodos são aplicáveis se o tamanho da amostra for grande o bastante. Apresentamos alguns conselhos práticos referentes à distribuição populacional e ao tamanho da amostra no fim desta seção.

Teste Unicaudal

Os **testes unicaudais** sobre a média de uma população assumem uma das duas seguintes formas:

Teste da Cauda Inferior	**Teste da Cauda Superior**
H_0: $\mu \geq \mu_0$	H_0: $\mu \leq \mu_0$
H_a: $\mu < \mu_0$	H_a: $\mu > \mu_0$

Consideremos um exemplo que envolve um teste da cauda inferior.

A Federal Trade Commission (FTC) realiza, periodicamente, estudos estatísticos concebidos para testar as afirmações feitas pelos fabricantes a respeito de seus produtos. Por exemplo, o rótulo de uma lata grande de Hilltop Coffee informa que a lata contém 3 libras (1,36 kg) de café. A FTC sabe que o processo de produção da Hilltop não consegue colocar exatamente 3 libras de café em cada lata, mesmo que o peso médio de enchimento da população de todas as latas cheias seja de, no mínimo, 3 libras por lata. Porém, contanto que o peso médio populacional seja de, no mínimo, 3 libras por lata, os direitos dos consumidores estarão garantidos. Desse modo, a FTC interpreta a informação contida no rótulo de uma lata grande de café como uma afirmação da parte da empresa Hilltop de que o peso médio populacional de enchimento é de, no mínimo, 3 libras por lata. Mostraremos como a FTC pode checar a afirmação da Hilltop realizando um teste de hipóteses da cauda inferior.

A primeira etapa consiste em desenvolver as hipóteses nula e alternativa para o teste. Se o peso médio de enchimento da população for, no mínimo, 3 libras por lata, a afirmação da Hilltop está correta. Esse resultado estabelece a hipótese nula para o teste. Entretanto, se o peso médio da população for inferior a 3 libras por lata, a afirmação da Hilltop está incorreta. Esse resultado estabelece a hipótese alternativa. Com μ denotando o peso médio de enchimento da população, as hipóteses nula e alternativa são as seguintes:

$$H_0: \mu \geq 3$$
$$H_a: \mu < 3$$

Observe que o valor hipotético da média populacional é $\mu_0 = 3$.

Se os dados amostrais indicarem que H_0 não pode ser rejeitada, as evidências estatísticas não sustentarão a conclusão de que ocorreu uma informação falsa no rótulo. Portanto, nenhuma ação deve ser praticada contra a Hilltop. No entanto, se os dados amostrais indicarem que H_0 pode ser rejeitada, concluiremos que a hipótese alternativa, H_a: $\mu < 3$, é verdadeira. Nesse caso, a conclusão de que há um volume menor de envasilhamento e uma acusação de informação falsa no rótulo se justificariam contra a Hilltop.

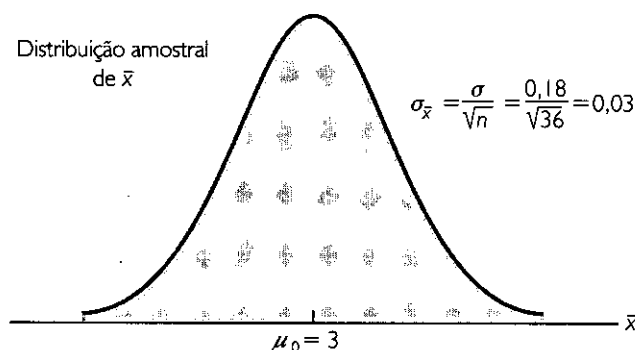
Suponha que uma amostra de 36 latas de café seja selecionada e que a média amostral $\bar{x}$ seja calculada como uma estimativa da média μ da população. Se o valor da média populacional $\bar{x}$ for inferior a 3 libras, os resultados da amostra lançarão dúvidas sobre a hipótese nula. O que queremos saber é a quantidade que $\bar{x}$ deve ser menor que 3 libras para nos dispormos a declarar que a diferença é significativa e arriscar-nos a cometer um erro do Tipo I ao acusar indevidamente a Hilltop de dar informações falsas no rótulo do produto. Um fator fundamental quando se trata dessa questão é o valor que o tomador de decisão seleciona para o nível de significância.

Conforme observamos na seção anterior, o nível de significância, denotado por α, é a probabilidade de se cometer um erro do Tipo I ao rejeitar H_0 quando a hipótese nula é verdadeira enquanto igualdade. O tomador de decisão deve especificar o nível de significância. Se o custo de cometer um erro do Tipo I for

elevado, um valor pequeno deve ser escolhido para o nível de significância. Se o custo não for elevado, um valor maior é mais apropriado. No estudo do Hilltop Coffee, o diretor do programa de testes da FTC fez a seguinte afirmação: "Se a empresa está cumprindo suas especificações de peso, com $\mu = 3$, não quero mover nenhum processo contra eles. Não obstante, estou disposto a arriscar uma chance de 1% de cometer esse erro". Em função da afirmação do diretor, definimos o nível de significância para o teste da hipótese em $\alpha = 0{,}01$. Desse modo, devemos projetar o teste da hipótese de forma que a probabilidade de cometermos um erro do Tipo I quando $\mu = 3$ seja 0,01.

Quanto ao estudo do Hilltop Coffee, ao desenvolvermos as hipóteses nula e alternativa e especificarmos o nível de significância para o teste, executamos as duas primeiras etapas necessárias à realização de todo teste de hipóteses. Agora, estamos preparados para executar a terceira etapa do teste de hipóteses: coletar os dados amostrais e calcular o valor daquilo que se denomina estatística de teste.

Estatística de teste Em relação ao estudo do Hilltop Coffee, testes anteriores realizados pela FTC mostram que o desvio padrão da população pode ser considerado conhecido, sendo o valor de $\sigma = 0{,}18$. Além disso, esses testes também mostram que se pode supor que a população de pesos de enchimento tenha uma distribuição normal. Em razão do estudo das distribuições amostrais no Capítulo 7, sabemos que se a população da qual extraímos a amostra está normalmente distribuída, a distribuição amostral de $\bar{x}$ também estará normalmente distribuída. Assim, para o estudo do Hilltop Coffee, a distribuição amostral de $\bar{x}$ está normalmente distribuída. Com um valor conhecido de $\sigma = 0{,}18$ e o tamanho amostral $n = 36$, a Figura 9.1 apresenta a distribuição amostral de $\bar{x}$ quando a hipótese nula é verdadeira enquanto igualdade; ou seja, quando $\mu = \mu_0 = 3$.* Note que o desvio padrão de $\bar{x}$ é dado por $\sigma_{\bar{x}} = \sigma/\sqrt{n} = 0{,}18/\sqrt{36} = 0{,}03$.

Figura 9.1 Distribuição amostral de $\bar{x}$ no estudo do Hilltop Coffee quando a hipótese nula é verdadeira enquanto igualdade ($\mu = \mu_0 = 3$)

Uma vez que a distribuição amostral de $\bar{x}$ está normalmente distribuída, a distribuição amostral de

$$z = \frac{\bar{x} - \mu_0}{\sigma_{\bar{x}}} = \frac{\bar{x} - 3}{0{,}03}$$

é uma distribuição normal padrão. Um valor de $z = -1$ significa que o valor de $\bar{x}$ está um erro padrão abaixo do valor hipotético da média, $z = -2$ significa que o valor de $\bar{x}$ está dois erros padrão abaixo do valor hipotético da média e assim por diante. Podemos usar a tabela de distribuição normal padrão para encontrar a probabilidade da cauda inferior correspondente a qualquer valor z. Por exemplo, a tabela normal padrão mostra que a área entre a média e $z = -3{,}00$ é 0,4987. Portanto, a probabilidade de se obter um valor de z que esteja três ou mais desvios padrão abaixo da média é $0{,}5000 - 0{,}4987 = 0{,}0013$. Em conseqüência, a probabilidade de se obter um valor de $\bar{x}$ que esteja três ou mais erros padrão abaixo da média populacional hipotética $\mu_0 = 3$ também é 0,0013. Esse resultado é improvável se a hipótese nula for verdadeira.

* Ao construir distribuições amostrais para testes de hipótese, presume-se que H_0 seja satisfeita enquanto igualdade.

Quanto aos testes de hipóteses sobre a média de uma população para o caso em que σ é desconhecido, usamos a variável aleatória z normal padrão como **estatística de teste** para determinar se $\bar{x}$ se desvia do valor hipotético μ o suficiente para justificar a rejeição da hipótese nula. Com $\sigma_{\bar{x}} = \sigma/\sqrt{n}$, a estatística de teste utilizada no caso em que σ é conhecido é a seguinte:

O desvio padrão de $\bar{x}$ é o desvio padrão da distribuição amostral de $\bar{x}$.

ESTATÍSTICA DE TESTE PARA TESTES DE HIPÓTESE A RESPEITO DE UMA MÉDIA POPULACIONAL: σ CONHECIDO

$$z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} \tag{9.1}$$

A questão fundamental relativa a um teste da cauda inferior é: quão pequena deve ser a estatística de teste z antes de optarmos por rejeitar a hipótese nula? Dois critérios podem ser utilizados para respondermos a essa questão.

O primeiro critério usa o valor z da estatística de teste para calcular uma probabilidade denominada **valor *p***. O valor p mede o suporte (ou a falta de suporte) que uma amostra dá à hipótese nula, e é a base para determinarmos se a hipótese nula deve ser rejeitada, dado o nível de significância. O segundo critério exige determinarmos em primeiro lugar um valor para a estatística de teste, chamado **valor crítico**. Para um teste da cauda inferior, o valor crítico vale como um ponto de referência para determinar se o valor da estatística de teste é pequeno o bastante para rejeitarmos a hipótese nula. Iniciamos com o critério do valor p.

Critério do valor *p* Na prática, o critério do valor p tornou-se o método preferível para determinar se a hipótese nula pode ser rejeitada, especialmente quando se usam softwares como o Minitab e o Excel. Para iniciar nossa discussão do uso dos valores p no teste de hipóteses, apresentamos agora uma definição formal de um valor p.

VALOR ***p***

O valor p é uma probabilidade, calculada usando-se a estatística de teste, que mede o apoio (ou a falta de apoio) proporcionado pela amostra à hipótese nula.

Visto que o valor p é uma probabilidade, ele varia de 0 a 1. Em geral, quanto maior o valor p, mais suporte a estatística de teste dá à hipótese nula. No entanto, um valor p pequeno indica uma estatística de teste da amostra que é incomum, dada a suposição de que H_0 é verdadeira. Valores p pequenos levam à rejeição de H_0, ao passo que valores p grandes indicam que a hipótese nula não deveria ser rejeitada.

Duas etapas são necessárias para usarmos o critério do valor p. Primeiro, devemos usar o valor da estatística de teste para calcular o valor p. O método usado para calcular o valor p depende de o teste ser da cauda inferior, da cauda superior ou bicaudal. Em relação a um teste da cauda inferior, o valor p é a probabilidade de obtermos um valor para a estatística de teste tão pequeno ou menor que aquele produzido pela amostra. Desse modo, para calcular o valor p relativo ao teste da cauda inferior no caso em que σ é conhecido, devemos encontrar a área sob a curva normal padrão à esquerda da estatística de teste. Depois de calcular o valor p, precisamos então decidir se ele é pequeno o bastante para rejeitar a hipótese nula; conforme veremos, essa decisão envolve comparar o valor p com o nível de significância.

ARQUIVO DA INTERNET
Coffee

Agora, vamos ilustrar o critério do valor p calculando o valor p do teste da cauda inferior para o Hilltop Coffee. Suponha que a amostra de 36 latas de café Hilltop produza uma média amostral $\bar{x} = 2{,}92$. Seria $\bar{x} = 2{,}92$ pequena o bastante para nos fazer rejeitar H_0? Desde que se trate de um teste da cauda inferior, o valor p é a área sob a curva normal padrão à esquerda da estatística de teste. Usando $\bar{x} = 2{,}92$, $\sigma = 0{,}18$ e $n = 36$, calculamos o valor z da estatística de teste.

$$z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} = \frac{2{,}92 - 3}{0{,}18/\sqrt{36}} = -2{,}67$$

Dessa forma, o valor p é a probabilidade de a estatística de teste z ser menor ou igual a –2,67 (a área sob a curva normal padrão à esquerda da estatística de teste).

Usando a tabela de distribuição normal padrão, descobrimos que a área entre a média e $z = -2{,}67$ é 0,4962. Assim, o valor p é 0,5000 – 0,4962 = 0,0038. A Figura 9.2 mostra que $\bar{x} = 2{,}92$ corresponde a $z = -2{,}67$ e a um valor $p = 0{,}0038$. Esse valor p indica uma pequena probabilidade de se obter uma média amostral $\bar{x} = 2{,}92$ (e uma estatística de teste igual a –2,67) ou menor quando se extrai a amostra de uma

população com $\mu = 3$. Esse valor p não dá um apoio muito consistente à hipótese nula, mas ele é pequeno o bastante para nos fazer rejeitar H_0? A resposta depende do nível de significância do teste.

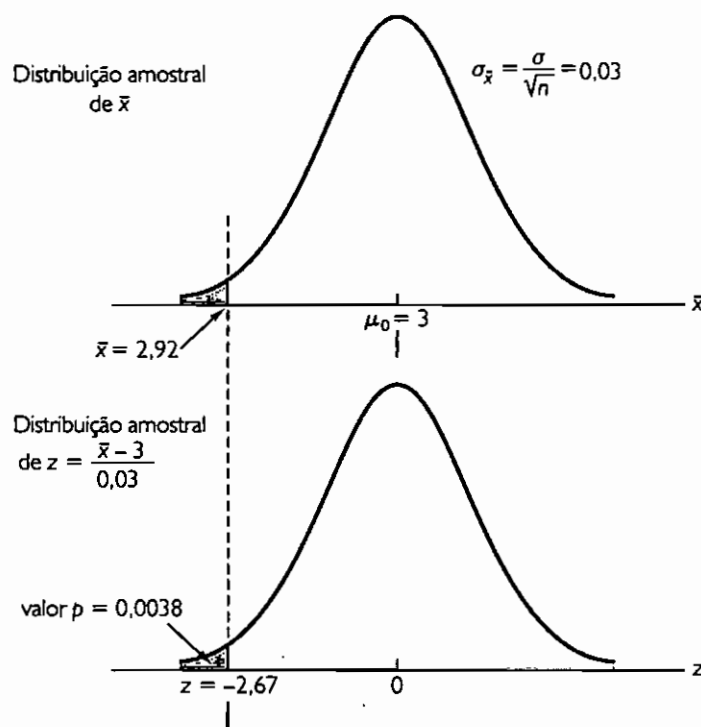
Conforme observamos anteriormente, o diretor do programa de testes da FTC selecionou um valor igual a 0,01 para o nível de significância. A escolha de $\alpha = 0{,}01$ significa que o diretor está disposto a aceitar uma probabilidade de 0,01 de rejeitar a hipótese nula quando ela for verdadeira enquanto igualdade ($\mu_0 = 3$). A amostra de 36 latas de café no estudo do Hilltop Coffee resultou em um valor $p = 0{,}0038$, o que significa que a probabilidade de se obter um valor $\bar{x} = 2{,}92$ ou menor quando a hipótese nula for verdadeira enquanto igualdade é 0,0038. Uma vez que 0,0038 é menor ou igual a $\alpha = 0{,}01$, rejeitamos H_0. Portanto, encontramos suficientes evidências estatísticas para rejeitar a hipótese nula dado o nível de significância de 0,01.

Agora, podemos formular a regra para determinar se a hipótese nula pode ser rejeitada quando se usa o critério do valor p. Para um nível de significância α, a regra de rejeição, quando se usa o critério do valor p, é a seguinte:

REGRA DE REJEIÇÃO QUANDO SE USA O VALOR p

Rejeitar H_0 se o valor $p \leq \alpha$

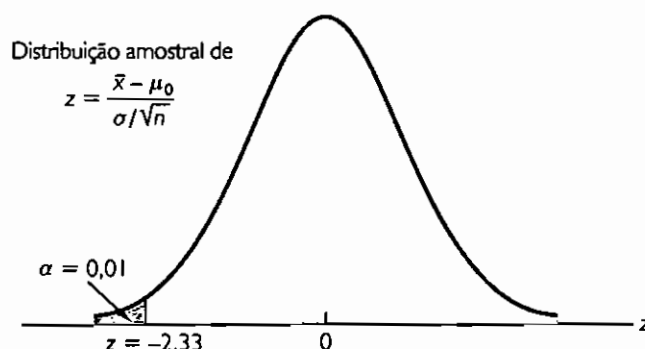
No teste do Hilltop Coffee, o valor p igual a 0,0038 resultou na rejeição da hipótese nula. Embora o fundamento para tomar a decisão de rejeitar envolva uma comparação do valor p com o nível de significância especificada pelo diretor da FTC, o valor p observado de 0,0038 significa que rejeitaríamos H_0 para qualquer valor $a \geq 0{,}0038$. Por esse motivo, o valor p também é chamado *nível observado de significância*.

Figura 9.2 Valor p para o estudo do Hilltop Coffee quando $\bar{x} = 2{,}92$ e $z = -2{,}67$

Diferentes tomadores de decisão podem exprimir diferentes opiniões quanto ao custo de cometer um erro do Tipo I e podem escolher um nível de significância diferente. Ao apresentar o valor p como parte dos resultados de testes de hipótese, outro tomador de decisão pode comparar o valor p relatado com o seu próprio nível de significância e tomar uma decisão diferente no que diz respeito a rejeitar H_0.

Critério do valor crítico Para um teste da cauda inferior, o valor crítico é o valor da estatística de teste que corresponde a uma área de α (o nível de significância) localizada na cauda inferior da distribuição amostral da estatística de teste. Em outras palavras, o valor crítico é o maior valor da estatística de teste que resultará na rejeição da hipótese nula. Vamos retornar ao exemplo do Hilltop Coffee e verificar como funciona essa abordagem.

No caso em que σ é conhecido, a distribuição amostral z da estatística de teste é uma distribuição normal padrão. Portanto, o valor crítico é o valor da estatística de teste que corresponde a uma área $\alpha = 0{,}01$ na cauda inferior de uma distribuição normal padrão. Usando a tabela de distribuição normal padrão, descobrimos que $z = -2{,}33$ produz uma área igual a 0,01 na cauda inferior (veja a Figura 9.3). Desse modo, se a amostra resultar em um valor da estatística de teste que seja menor ou igual a –2,33, o valor p correspondente será menor ou igual a 0,01; nesse caso, deveríamos rejeitar a hipótese nula. Portanto, para o estudo do Hilltop Coffee, a regra de rejeição pelo critério do valor crítico com um nível de significância de 0,01 é

$$\text{Rejeitar } H_0 \text{ se } z \leq -2{,}33$$

Figura 9.3 Valor crítico = –2,33 para o teste de hipóteses do Hilltop Coffee

No exemplo do Hilltop Coffee, $\bar{x} = 2{,}92$ e a estatística de teste é $z = -2{,}67$. Uma vez que $z = -2{,}67 < -2{,}33$, podemos rejeitar H_0 e concluir que a empresa Hilltop Coffee está preenchendo as latas com um volume menor.

Podemos generalizar a regra de rejeição pelo critério do valor crítico para manipular qualquer nível de significância. A regra de rejeição para um teste da cauda inferior é a seguinte:

REGRA DE REJEIÇÃO PARA UM TESTE DA CAUDA INFERIOR: CRITÉRIO DO VALOR CRÍTICO

$$\text{Rejeitar } H_0 \text{ se } z \leq -z_\alpha$$

em que $-z_\alpha$ é o valor crítico; ou seja, o valor que produz uma área α na cauda inferior da distribuição normal padrão.

O critério do valor p para testes de hipótese e o critério do valor crítico sempre levarão à mesma decisão de rejeição; ou seja, quando se quer que o valor p seja menor ou igual a α, o valor da estatística de teste será menor ou igual ao valor crítico. A vantagem do critério do valor p é que o valor p nos diz *quão* significativos são os resultados (o nível observado de significância). Quando usamos o critério do valor crítico, sabemos que os resultados são significativos ao nível declarado de significância.

Há procedimentos computadorizados de teste de hipóteses que fornecem o valor p, de forma que este está se tornando rapidamente o método preferido de realizar testes de hipóteses. Se não tiver acesso a um computador, talvez você prefira usar o critério do valor crítico. Para algumas distribuições de probabilidade é mais fácil usar tabelas estatísticas para encontrar um valor crítico do que usar as tabelas para calcular o valor p. Esse tópico será discutido com mais detalhes na próxima seção.

No início desta seção, dissemos que os testes unicaudais a respeito de uma média populacional assumem uma das duas seguintes formas:

Teste da Cauda Inferior	**Teste da Cauda Superior**
$H_0: \mu \geq \mu_0$	$H_0: \mu \leq \mu_0$
$H_a: \mu < \mu_0$	$H_a: \mu > \mu_0$

Usamos o estudo do Hilltop Coffee para ilustrar como realizar um teste da cauda inferior. Podemos usar o mesmo critério geral para realizar um teste da cauda superior. A estatística de teste z ainda é calculada usando-se a Equação 9.1. Porém, para um teste da cauda superior, o valor p é a probabilidade de obter um valor para a estatística de teste que seja tão grande ou maior que aquele que é produzido pela amostra.

Desse modo, para calcular um valor p para o teste da cauda superior no caso em que σ é conhecido, devemos encontrar uma área sob a curva normal padrão à direita da estatística de teste. O uso do critério do valor crítico faz que rejeitemos a hipótese nula se o valor da estatística de teste for maior ou igual ao valor crítico z_α; em outras palavras, rejeitamos H_0 se $z \geq z_\alpha$.

Teste Bicaudal

Nos testes de hipótese, a regra para um **teste bicaudal** a respeito de uma média populacional é expressa da seguinte maneira:

$$H_0: \mu = \mu_0$$
$$H_a: \mu \neq \mu_0$$

Nesta subseção, mostramos como realizar um teste bicaudal a respeito de uma média populacional para o caso em que σ é conhecido. Como ilustração, considere a situação de teste de hipóteses enfrentada pela MaxFlight, Inc.

A U.S. Golf Association (USGA) estabelece normas que os fabricantes de equipamentos de golfe devem cumprir para que seus produtos aceitos e usados nos eventos da USGA. A MaxFlight utiliza um processo de manufatura de alta tecnologia para produzir bolas de golfe que atingem uma distância média de arremesso (*driving distance*) de 295 jardas (269,7 m). Às vezes, porém, o processo se desajusta e produz bolas de golfe que atingem uma distância média de arremesso diferente de 295 jardas. Quando a distância média cai abaixo de 295 jardas, a empresa se preocupa em perder vendas pelo fato de as bolas de golfe não atingirem a distância anunciada. Quando a distância média passa de 295 jardas, as bolas de golfe da MaxFlight podem ser rejeitadas pela USGA em virtude de excederem o padrão de distância total referente ao *carry and roll*.[1]

O programa de controle da qualidade da MaxFlight envolve extrair amostras periódicas de 50 bolas de golfe para monitorar o processo de manufatura. Para cada amostra, é realizado um teste de hipóteses com o objetivo de determinar se o processo se desajustou. Vamos desenvolver as hipóteses nula e alternativa. Iniciamos, supondo que o processo esteja funcionando corretamente; ou seja, as bolas de golfe que são produzidas atingem uma distância média de 295 jardas. Essa suposição estabelece a hipótese nula. A hipótese alternativa é que a distância média não é igual a 295 jardas. Com um valor hipotético de $\mu_0 = 295$, as hipóteses nula e alternativa do teste de hipóteses da MaxFlight são as seguintes:

$$H_0: \mu = 295$$
$$H_a: \mu \neq 295$$

Se a média amostral $\bar{x}$ for significativamente menor que 295 jardas ou significativamente maior que 295 jardas, rejeitaremos H_0. Nesse caso, serão tomadas medidas corretivas para ajustar o processo de manufatura. No entanto, se $\bar{x}$ não se desviar da média hipotética $\mu_0 = 295$ em termos de um valor significativo, H_0 não será rejeitada e nenhuma ação será encaminhada para ajustar o processo de manufatura.

A equipe de controle da qualidade selecionou $\alpha = 0{,}05$ como o nível de significância para o teste. Dados de testes anteriores, realizados quando se sabia que o processo estava devidamente ajustado, mostram que se pode presumir que o desvio padrão da população seja conhecido, tendo o valor $\sigma = 12$. Desse modo, com um tamanho de amostra $n = 50$, o desvio padrão de $\bar{x}$ é:

[1] NT: *Carry and roll* – Muitas tacadas de golfe fazem a bola viajar pelo ar (*carry*) e rolar (*roll*) certa distância. A distância total percorrida pela bola nesse processo denomina-se *carry and roll* (Golfe).

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{12}{\sqrt{50}} = 1,7$$

Uma vez que o tamanho da amostra é grande, o teorema do limite central (veja o Capítulo 7) nos permite concluir que a distribuição amostral de $\bar{x}$ pode ser aproximada por uma distribuição normal.

A Figura 9.4 apresenta a distribuição amostral de $\bar{x}$ referente ao teste de hipóteses da MaxFlight, considerando uma média populacional hipotética de $\mu_0 = 295$.

Suponha que uma amostra de 50 bolas de golfe seja selecionada e que a média da amostra seja $\bar{x} = 297,6$ jardas. Essa média amostral sustenta a conclusão de que a média populacional é maior que 295 jardas. Esse valor de $\bar{x}$ é suficientemente maior que 295 para nos fazer rejeitar H_0 ao nível de significância 0,05? Na seção anterior, descrevemos dois critérios que podem ser usados para responder a essa pergunta: o critério do valor p e o critério do valor crítico.

Critério do valor *p* Lembre-se de que o valor p é uma probabilidade, calculada usando-se a estatística de teste, para medir o apoio (ou a falta de apoio) que a amostra dá à hipótese nula. Em um teste bicaudal, valores da estatística de teste que se encontram em *qualquer uma* das caudas indicam falta de suporte à hipótese nula. Em um teste bicaudal, o valor p é a probabilidade de se obter um valor para a estatística de teste *tão ou mais improvável* do que aquele que é fornecido pela amostra. Vejamos como o valor p é calculado para o teste de hipóteses da MaxFlight.

Primeiramente, calculamos o valor da estatística de teste. Para o caso em que σ é conhecido, a estatística de teste z é uma variável aleatória normal padrão. Usando a Equação 9.1 com $\bar{x} = 297,6$, o valor da estatística de teste é:

$$z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} = \frac{297,6 - 295}{12/\sqrt{50}} = 1,53$$

Agora, para calcular o valor p, devemos encontrar a probabilidade de obtermos um valor para a estatística de teste que seja, *no mínimo, tão improvável quanto* $z = 1,53$. Evidentemente, valores de $z \geq 1,53$ são, *no mínimo, tão improváveis* quanto esse valor. Porém, já que este é um teste bicaudal, valores de $z \leq -1,53$ também são, *no mínimo, tão improváveis* quanto o valor da estatística de teste fornecido pela amostra. Consultando a Figura 9.5, notamos que o valor p bicaudal, nesse caso, é dado por $P(z \leq -1,53) + P(z \leq 1,53)$. Uma vez que a curva normal é simétrica, podemos calcular essa probabilidade encontrando a área sob a curva normal padrão à direita de $z = 1,53$ e a duplicando. A tabela da distribuição normal padrão mostra que a área entre a média e $z = 1,53$ é 0,4370. Assim, a área sob a curva normal padrão à direita da estatística de teste $z = 1,53$ é $0,5000 - 0,4370 = 0,0630$. Duplicando esse valor, descobrimos que o valor p para o teste de hipótese bicaudal da MaxFlight é valor $p = 2(0,0630) = 0,1260$.

Em seguida, comparamos o valor p com o nível de significância para verificar se a hipótese nula deveria ser rejeitada. Com um nível de significância $\alpha = 0,05$, não rejeitamos H_0 porque o valor $p = 0,1260 > 0,05$. Desde que a hipótese nula não seja rejeitada, nenhuma ação será tomada para ajustar o processo de manufatura da MaxFlight.

Figura 9.4 Distribuição amostral de $\bar{x}$ para o teste de hipóteses da MaxFlight

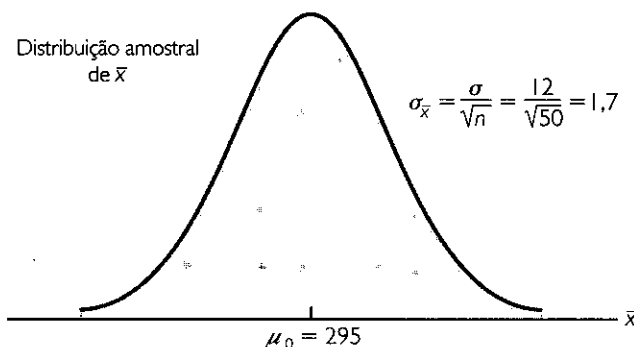
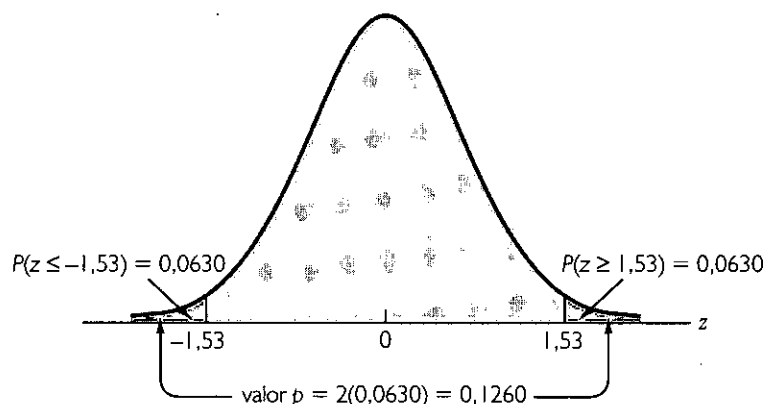



Figura 9.5 Valor p do teste de hipóteses da MaxFlight

O cálculo do valor p de um teste bicaudal pode parecer um pouco confuso em comparação com o cálculo do valor p de um teste unicaudal. Entretanto, ele pode ser simplificado pelas três etapas seguintes:

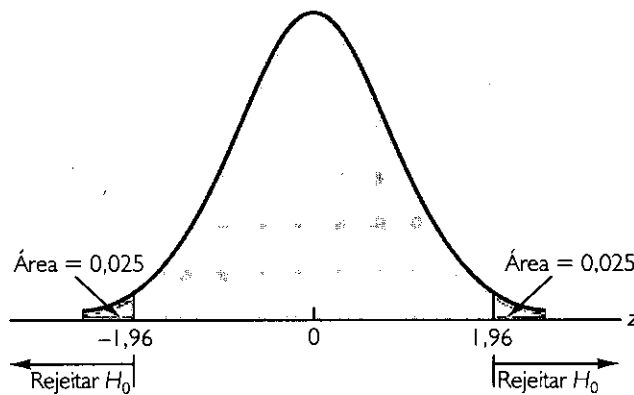
1. Calcule o valor da estatística de teste z.
2. Se o valor da estatística de teste estiver na cauda superior ($z > 0$), encontre a área sob a curva normal padrão à direita de z. Se o valor da estatística de teste estiver na cauda inferior, encontre a área da curva normal padrão à esquerda de z.
3. Duplique a área da cauda, ou probabilidade, obtida na etapa 2 para obter o valor p.

Na prática, o cálculo do valor p é feito automaticamente quando se usa softwares como o Minitab ou o Excel. Por exemplo, a Figura 9.6 mostra a saída de dados (*output*) do Minitab relativa ao teste de hipóteses da MaxFlight. A média amostral $\bar{x} = 297{,}6$, a estatística de teste $z = 1{,}53$ e o valor $p = 0{,}126$ estão em destaque. O procedimento passo a passo para obter a saída de dados do Minitab é descrito no Apêndice 9.1.

Critério do valor crítico Antes de sairmos desta seção, vejamos como a estatística de teste z pode ser comparada com um valor crítico para se tomar a decisão do teste de hipóteses referente a um teste bicaudal. A Figura 9.7 indica que os valores críticos do teste ocorrerão tanto na cauda inferior quanto na cauda superior da distribuição normal padrão. Com um nível de significância $\alpha = 0{,}05$, a área em cada cauda, além dos valores críticos, é $\alpha/2 = 0{,}05/2 = 0{,}025$. Usando a tabela de áreas da distribuição normal padrão, descobrimos que os valores críticos da estatística de teste são $-z_{0,025} = -1{,}96$ e $z_{0,025} = 1{,}96$.

Figura 9.6 Saída de dados do Minitab relativa ao teste de hipóteses da MaxFlight

```
Test of mu = 295 vs not = 295
The assumed sigma = 12

Variable     N      Mean      StDev    SE Mean
Yards       50   297.600     11.297      1.697

Variable          95.0% CI             Z        P
Yards       (294.274, 300.926)      1.53    0.126
```

Figura 9.7 Valores críticos relativos ao teste de hipóteses da MaxFlight

Desse modo, usando-se o critério do valor crítico, a regra de rejeição bicaudal é:

$$\text{Rejeitar } H_0 \text{ se } z \leq -1{,}96 \text{ ou se } z \geq 1{,}96$$

Uma vez que o valor da estatística de teste do estudo da MaxFlight é $z = 1{,}53$, a evidência estatística não nos permitirá rejeitar a hipótese nula ao nível de significância 0,05.

Resumo e Conselho Prático

Apresentamos exemplos de teste da cauda inferior e da cauda superior a respeito de uma média populacional. Baseando-se nesses exemplos, agora podemos resumir os procedimentos de teste de hipóteses a respeito de uma média populacional para o caso em que σ é conhecido, como mostra a Tabela 9.2. Observe que μ_0 é o valor hipotético da média populacional.

As etapas de teste de hipóteses seguidas nos dois exemplos exibidos nesta seção são comuns a todo teste de hipóteses.

ETAPAS DO TESTE DE HIPÓTESES

Etapa 1. Desenvolver as hipóteses nula e alternativa.
Etapa 2. Especificar o nível de significância.
Etapa 3. Coletar os dados da amostra e calcular o valor da estatística de teste.

Critério do valor p

Etapa 4. Usar o valor da estatística de teste para calcular o valor p.
Etapa 5. Rejeitar H_0 se o valor $p \leq \alpha$.

Critério do valor crítico

Etapa 4. Usar o nível de significância para estabelecer o valor crítico e o valor de rejeição.
Etapa 5. Usar o valor da estatística de teste e a regra de rejeição para determinar se é oportuno rejeitar H_0.

Tabela 9.2 Resumo dos testes de hipótese a respeito de uma média populacional: caso em que σ é conhecido

	Teste da Cauda Inferior	Teste da Cauda Superior	Teste Bicaudal
Hipótese	H_0: $\mu \geq \mu_0$ H_a: $\mu < \mu_0$	H_0: $\mu \leq \mu_0$ H_a: $\mu > \mu_0$	H_0: $\mu = \mu_0$ H_a: $\mu \neq \mu_0$
Estatística de Teste	$z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$	$z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$	$z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$
Regra de Rejeição: Critério do Valor *p*	Rejeitar H_0 se o valor $p \leq \alpha$	Rejeitar H_0 se o valor $p \leq \alpha$	Rejeitar H_0 se o valor $p \leq \alpha$
Regra de Rejeição: Critério do Valor Crítico	Rejeitar H_0 se $z \leq -z_\alpha$	Rejeitar H_0 se $z \geq z_\alpha$	Rejeitar H_0 se $z \leq -z_{\alpha/2}$ ou se $z \geq z_{\alpha/2}$

O conselho prático sobre o tamanho da amostra para testes de hipótese é idêntico àquele que apresentamos acerca do tamanho da amostra para estimação de intervalos no Capítulo 8. Na maioria das aplicações, um tamanho de amostra $n \geq 30$ é adequado quando se usa o procedimento de teste de hipóteses descrito nesta seção. Nos casos em que o tamanho da amostra é inferior a 30, a distribuição da população da qual extraímos a amostra torna-se um fator importante. Se a população está normalmente distribuída, o procedimento de teste de hipóteses que acabamos de descrever é exato e pode ser usado para qualquer tamanho de amostra. Se a população não está normalmente distribuída, mas é pelo menos aproximadamente simétrica, pode-se esperar que tamanhos de amostras pequenos, até mesmo iguais a 15, produzam resultados aceitáveis. Com tamanhos de amostra menores, o procedimento desse teste de hipóteses mostrado nesta seção somente será usado se o analista acreditar, ou estiver disposto a assumir, que a população está pelo menos aproximadamente distribuída.

Relação entre a Estimação por Intervalo e o Teste de Hipóteses

Encerramos esta seção discutindo a relação entre a estimação por intervalo e o teste de hipóteses. No Capítulo 8, mostramos como desenvolver uma estimação por intervalo de confiança de uma média populacional. Para o caso em que *s* é conhecido, a estimação por intervalo de confiança de uma média populacional correspondente a um coeficiente de confiança $1 - a$ é dada por:

$$\bar{x} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \tag{9.2}$$

A realização de um teste de hipóteses requer que desenvolvamos primeiro as hipóteses a respeito do valor de um parâmetro populacional. No caso de uma média populacional, o teste bicaudal assume a forma:

$$H_0: \mu = \mu_0$$
$$H_a: \mu \neq \mu_0$$

em que μ_0 é o valor hipotético da média da população. Utilizando o critério do valor crítico bicaudal, não rejeitamos H_0 para valores da média amostral $\bar{x}$ que estão dentro dos intervalos de erro padrão $-z_{\alpha/2}$ e $z_{\alpha/2}$ de μ_0. Desse modo, a região "não rejeitar" da média amostral $\bar{x}$ em um teste de hipóteses bicaudal com um nível de significância α é dada por:

$$\mu_0 \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \tag{9.3}$$

Um exame mais cuidadoso das Equações 9.2 e 9.3 fornece subsídios para que se possa compreender a relação entre os critérios de estimação e testes de hipóteses com a inferência estatística. Observe, em especial, que ambos os procedimentos requerem o cálculo dos valores $z_{\alpha/2}$ e $\sigma/\sqrt{n}$. Concentrando-se em α, notamos que o coeficiente de confiança $(1 - \alpha)$ da estimação por intervalo corresponde a um nível de significância α no teste de hipóteses. Por exemplo, um intervalo de confiança de 95% corresponde a um nível de significância de 0,05 para o teste de hipóteses. Além disso, as Equações 9.2 e 9.3 mostram que, desde

que $z_{\alpha/2}$ $(\sigma/\sqrt{n})$ é o valor positivo (mais) ou negativo (menos) de ambas as expressões, se $\bar{x}$ estiver na região "não rejeitar" definida pela Equação 9.3, o valor hipotético μ_0 estará no intervalo de confiança definido pela Equação 9.2. Inversamente, se o valor hipotético μ_0 estiver no intervalo de confiança definido pela Equação 9.2, a média amostral $\bar{x}$ estará na região "não rejeitar" da hipótese H_0: $\mu - \mu_0$, conforme é definido pela Equação 9.3. Essas observações acarretam o seguinte procedimento para se usar um intervalo de confiança a fim de realizar um teste bicaudal.

CRITÉRIO DO INTERVALO DE CONFIANÇA PARA TESTAR UMA HIPÓTESE DA FORMA

$$H_0: \mu = \mu_0$$
$$H_a: \mu \neq \mu_0$$

1. Selecione uma amostra aleatória simples da população e use o valor da média amostral $\bar{x}$ para desenvolver o intervalo de confiança da média populacional *m*.

$$\bar{x} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$

2. Se o intervalo de confiança contiver o valor hipotético μ_0, não rejeite H_0. Caso contrário, rejeite H_0.

Para um teste de hipótese bicaudal, a hipótese nula pode ser rejeitada se o intervalo de confiança não incluir μ_0.

Retornemos ao teste de hipóteses da MaxFlight, o qual resultou no seguinte teste bicaudal:

$$H_0: \mu = 295$$
$$H_a: \mu \neq 295$$

Para testar essa hipótese com um nível de significância $\alpha = 0{,}05$, extraímos uma amostra de 50 bolas de golfe e encontramos uma distância média amostral $\bar{x} = 297{,}6$ jardas. Lembre-se de que o desvio padrão populacional $\sigma = 12$. Usando esses resultados com $z_{0,025} = 1{,}96$, descobrimos que a estimação por intervalo de confiança de 95% da média populacional é:

$$\bar{x} \pm z_{0,025} \frac{\sigma}{\sqrt{n}}$$

$$297{,}6 \pm 1{,}96 \frac{12}{\sqrt{50}}$$

$$297{,}6 \pm 3{,}3$$

ou

$$294{,}3 \text{ a } 300{,}9$$

Esse resultado possibilita ao gerente de controle da qualidade concluir com 95% de confiança que a distância média atingida pela população das bolas de golfe está entre 294,3 e 300,9 jardas (269,10 m e 275,14 m, respectivamente). Uma vez que o valor hipotético da média populacional, $\mu_0 = 295$, está contido nesse intervalo, a conclusão do teste de hipóteses é que a hipótese nula, H_0: $\mu = 295$, não pode ser rejeitada.

Note que essa discussão e exemplo pertencem a testes de hipótese bicaudais a respeito de uma média populacional. Entretanto, existe a mesma relação entre o intervalo de confiança e os testes de hipótese bicaudais para outros parâmetros populacionais. A relação também pode ser estendida para testes unicaudais a respeito de parâmetros populacionais. Para fazê-lo, porém, é necessário o desenvolvimento de intervalos de confiança unilaterais, os quais raramente são usados na prática.

NOTAS E COMENTÁRIOS

1. No Apêndice 9.2, mostramos como calcular valores *p* com o Excel.
2. Quanto menor o valor *p*, maior a evidência contra H_0, bem como a favor de H_a. Eis algumas diretrizes estatísticas que os estatísticos sugerem para interpretar valores *p* pequenos:
   - Menor que 0,01 – Esmagadora evidência de que H_a é verdadeira.
   - Entre 0,01 e 0,05 – Forte evidência de que H_0 é verdadeira.

- Entre 0,05 e 0,10 – Fraca evidência de que H_a é verdadeira.
- Maior que 0,10 – Insuficiente evidência de que H_a é verdadeira.

Exercícios

Nota para o estudante: Alguns dos exercícios que são apresentados a seguir lhe pedem para usar o critério do valor p, e outros pedem para usar o critério do valor crítico. Ambos os métodos produzirão a mesma conclusão de um teste de hipóteses. Apresentamos exercícios com os dois métodos para lhe dar a oportunidade de praticar ao utilizar ambos. Nas seções posteriores e nos capítulos seguintes, geralmente enfatizaremos o critério do valor p como o método preferível, mas você pode escolher qualquer um dos dois baseando-se em sua preferência pessoal.

Métodos

9. Considere o seguinte teste de hipóteses:

$$H_0: \mu \geq 20$$
$$H_a: \mu < 20$$

Uma amostra de tamanho 50 produziu a média amostral 19,4. O desvio padrão da população é 2.

a. Calcule o valor da estatística de teste.
b. Qual é o valor p?
c. Usando $\alpha = 0{,}05$, qual é a sua conclusão?
d. Qual é a regra de rejeição, usando-se o valor crítico? Qual é a sua conclusão?

AUTOTESTE

10. Considere o seguinte teste de hipóteses:

$$H_0: \mu \leq 25$$
$$H_a: \mu > 25$$

Uma amostra de tamanho 40 produziu a média amostral 26,4. O desvio padrão da população é 6.

a. Calcule o valor da estatística de teste.
b. Qual é o valor p?
c. Com $\alpha = 0{,}01$, qual é a sua conclusão?
d. Qual é a regra de rejeição, usando-se o valor crítico?

AUTOTESTE

11. Considere o seguinte teste de hipóteses:

$$H_0: \mu = 15$$
$$H_a: \mu \neq 15$$

Uma amostra de tamanho 50 produziu a média amostral 14,15. O desvio padrão da população é 3.

a. Calcule o valor da estatística de teste.
b. Qual é o valor p?
c. Com $\alpha = 0{,}05$, qual é a sua conclusão?
d. Qual é a regra de rejeição, usando-se o valor crítico? Qual é a sua conclusão?

12. Considere o seguinte teste de hipóteses:

$$H_0: \mu \geq 80$$
$$H_a: \mu < 80$$

Uma amostra de tamanho 100 é usada e o desvio padrão da população é 12. Calcule o valor p e apresente sua conclusão quanto a cada um dos seguintes resultados amostrais. Use $\alpha = 0{,}01$.

a. $\bar{x} = 78{,}5$
b. $\bar{x} = 77$
c. $\bar{x} = 75{,}5$
d. $\bar{x} = 81$

13. Considere o seguinte teste de hipóteses:

$$H_0: \mu \leq 50$$
$$H_a: \mu > 5$$

Uma amostra de tamanho 60 é usada e o desvio padrão da população é 8. Use o critério do valor crítico para apresentar sua conclusão quanto a cada um dos seguintes resultados amostrais. Use $\alpha = 0{,}05$.

a. $\bar{x} = 52{,}5$
b. $\bar{x} = 51$
c. $\bar{x} = 51{,}8$

14. Considere o seguinte teste de hipóteses:

$$H_0: \mu = 22$$
$$H_a \mu \neq 22$$

Uma amostra de tamanho 75 é usada e o desvio padrão da população é 10. Calcule o valor p e apresente sua conclusão quanto a cada um dos seguintes dados amostrais. Use $\alpha = 0{,}01$.

a. $\bar{x} = 23$
b. $\bar{x} = 25{,}1$
c. $\bar{x} = 20$

Aplicações

AUTOTESTE

15. As declarações do imposto de renda individuais entregues antes do dia 31 de março obtiveram uma média de restituição de US$ 1.056. Considere a população de declarantes "de última hora" que entregam suas declarações durante os cinco últimos dias do período de entrega das declarações do imposto de renda (tipicamente, de 10 a 15 de abril).
   a. Um pesquisador sugere que uma razão para que as pessoas esperem até os cinco últimos dias é que em média essas pessoas têm menores restituições a receber do que aquelas que entregam as declarações primeiro. Desenvolva as hipóteses apropriadas de tal forma que a rejeição de H_0 sustente a argumentação do pesquisador.
   b. Para uma média de 400 indivíduos que entregaram suas declarações entre 10 e 15 de abril, a média amostral de restituição foi de US$ 910. Baseando-se na experiência anterior, pode-se supor um desvio padrão populacional σ = US$ 1.600. Qual é o valor p?
   c. Com $\alpha = 0{,}05$, qual é a sua conclusão?
   d. Repita o teste de hipóteses anterior usando o critério do valor crítico.

16. A Reis, Inc., uma firma de pesquisa imobiliária de Nova York, acompanha o custo do aluguel de apartamentos nos Estados Unidos. Em meados de 2002, o índice médio de aluguel por apartamento em todo o território nacional era de US$ 895 por mês (*The Wall Street Journal*, 8 de julho de 2002). Suponha que, baseando-se em pesquisas trimestrais históricas, seja razoável considerar-se um desvio padrão populacional σ = US$ 225. Em um estudo recente dos índices de aluguel de apartamentos, uma amostra de 180 apartamentos de todo o país produziu uma média amostral de US$ 915 por mês. Os dados amostrais possibilitam à Reis concluir que o índice médio populacional de aluguel de apartamentos agora ultrapasse o nível relatado em 2002?
   a. Estabeleça as hipóteses nula e alternativa.
   b. Qual é o valor p?
   c. Com $\alpha = 0{,}01$, qual é a sua conclusão?
   d. O que você recomendaria que a Reis considerasse fazer agora?

17. Foi divulgado que a duração média de uma semana de trabalho para a população de trabalhadores é de 39,2 horas (*Investor's Business Daily*, 11 de setembro de 2000). Suponha que quiséssemos extrair uma amostra atual de trabalhadores para verificar se a duração média de uma semana de trabalho se modificou das 39,2 horas relatadas anteriormente.
   a. Estabeleça as hipóteses que nos ajudem a determinar se ocorreu uma alteração na duração média da semana de trabalho.
   b. Suponha que um tamanho de amostra de 112 trabalhadores tenha produzido uma média amostral de 38,5 horas. Use um desvio padrão populacional $\sigma = 4{,}8$ horas. Qual é o valor p?
   c. Com $\alpha = 0{,}05$, a hipótese nula pode ser rejeitada? Qual é a sua conclusão?
   d. Repita o teste de hipótese anterior usando o critério do valor crítico.

18. A média de rendimento anual total dos fundos mútuos de ações diversificados – U.S. Diversified Equity funds – de 1999 a 2003 foi de 4,1% (*Business Week*, 26 de janeiro de 2004). Um pesquisador gostaria de realizar um teste de hipóteses para verificar se os rendimentos dos fundos de crescimento de média capitalização (*mid-cap growth funds*), ao longo do mesmo período, são significativamente diferentes da média dos fundos mútuos de ações diversificados.

a. Formule as hipóteses que podem ser usadas para determinar se a média de rendimento anual dos fundos de crescimento de média capitalização difere da média dos fundos mútuos de ações diversificados.
b. Uma amostra de 40 fundos de crescimento de média capitalização fornece uma média de retorno anual $\bar{x} = 3,4\%$. Suponha que se saiba, em decorrência dos estudos anteriores, que o desvio padrão da população dos fundos de crescimento de média capitalização seja $\sigma = 2\%$; use os resultados amostrais para calcular a estatística de teste e o valor p do teste de hipóteses.
c. Com $\alpha = 0,05$, qual é a sua conclusão?

19. Em 2001, o U.S. Department of Labor (Departamento do Trabalho dos Estados Unidos) relatou que a média de remuneração horária para os trabalhadores do setor de produção norte-americanos é de US$ 14,32 por hora (*The World Almanac*, 2003). Uma amostra de 75 trabalhadores do setor de produção durante 2003 produziu uma média amostral de US$ 14,68 por hora. Supondo que o desvio padrão da população seja $\sigma =$ US$ 1,45, podemos concluir que ocorreu um aumento da remuneração horária média a partir de 2001? Use $\alpha = 0,05$.

20. A média nacional dos preços de venda de casas novas destinadas a uma única família é US$ 181.900 (*The New York Times Almanac*, 2000). Uma amostra de 40 vendas de casas destinadas a uma única família no sul do país exibiu uma média amostral igual a US$ 166.400. Use o desvio padrão populacional de US$ 33.500.
a. Formule as hipóteses nula e alternativa que podem ser usadas para determinar se os dados amostrais sustentam a conclusão de que a média populacional dos preços de venda de casas novas destinadas a uma única família no sul do país seja menor que a média nacional de US$ 181.900.
b. Qual é o valor da estatística de teste?
c. Qual é o valor p?
d. Com $\alpha = 0,01$, qual é a sua conclusão?

21. A Fowle Marketing Research, Inc., fundamenta os preços que cobra de seus clientes na suposição de que as pesquisas telefônicas podem ser concluídas em um tempo médio de 15 minutos ou menos. Se for necessário um tempo médio de pesquisa mais longo, uma taxa adicional é cobrada. Suponha que uma amostra de 35 pesquisas apresente uma média amostral de 17 minutos. Use $\sigma = 4$ minutos. A taxa adicional se justifica?
a. Formule as hipóteses nula e alternativa para essa aplicação.
b. Calcule o valor da estatística de teste.
c. Qual é o valor p?
d. Com $\alpha = 0,01$, qual é a sua conclusão?

22. A CCN e a ActMedia criaram um canal de televisão destinado a pessoas que esperam nas filas do caixa de supermercados. O canal apresentava notícias, entrevistas breves e anúncios. A duração do programa baseava-se na suposição de que o tempo médio que a população de compradores permanece em uma fila de supermercado é igual a 8 minutos. Uma amostra de tempos de espera reais será usada para testar essa suposição e determinar se a média de tempo de espera real difere desse padrão.
a. Formule as hipóteses para essa aplicação.
b. Uma amostra de 120 compradores apresentou uma média amostral de tempo de espera de 8,5 minutos. Suponha um desvio padrão populacional $s = 3,2$ minutos. Qual é o valor p?
c. Com $\alpha = 0,05$, qual é a sua conclusão.
d. Calcule um intervalo de confiança de 95% para a média populacional. Ela sustenta sua conclusão?

9.4 MÉDIA DA POPULAÇÃO: σ DESCONHECIDO

Nesta seção, descreveremos como realizar testes de hipótese a respeito de uma média populacional considerando o caso em que σ é desconhecido. Uma vez que o caso em que σ é desconhecido corresponde à situação em que não se pode desenvolver uma estimativa do desvio padrão populacional antes de se fazer a amostragem, a amostra deve ser usada para desenvolver uma estimativa de μ, tanto quanto de σ. Assim, para se realizar um teste de hipóteses a respeito de uma média populacional para o caso em que σ é desconhecido, utilizamos a média amostral $\bar{x}$ como uma estimativa de μ e usamos o desvio padrão s da amostra como uma estimativa de σ.

As etapas do procedimento de teste de hipóteses referentes ao caso em que σ é desconhecido são similares às do caso em que σ é conhecido, conforme descrevemos na Seção 9.3. Mas, com σ desconhecido, os cálculos da estatística de teste e do valor p são bem diferentes. Lembre-se de que, no caso em que σ é conhecido, a distribuição amostral da estatística de teste tem uma distribuição normal padrão. Porém, para o caso em que σ é desconhecido, a distribuição amostral da estatística de teste tem uma variabilidade ligeiramente maior porque a amostra é usada para desenvolver estimativas tanto de μ como de σ.

Na Seção 8.2, mostramos que uma estimação por intervalo de uma média populacional para o caso em que σ é desconhecido baseia-se em uma distribuição probabilística conhecida por *distribuição t*. Os testes de hipóteses a respeito da média de uma população para o caso em que σ é desconhecido também se baseiam na distribuição t. Para o caso em que σ é desconhecido, a estatística de teste tem uma distribuição t com $n - 1$ graus de liberdade.

ESTATÍSTICA DE TESTE PARA TESTES DE HIPÓTESE A RESPEITO DE UMA MÉDIA POPULACIONAL: σ DESCONHECIDO

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} \tag{9.4}$$

No Capítulo 8, dissemos que a distribuição t se baseia na suposição de que a população da qual extraímos a amostra tem uma distribuição amostral. Entretanto, as pesquisas mostram que essa suposição pode ser consideravelmente desprezada quando o tamanho da amostra for suficientemente grande. Apresentamos alguns conselhos práticos referentes à distribuição populacional e ao tamanho da amostra no fim desta seção.

Teste Unicaudal

ARQUIVO DA INTERNET
AirRating

Consideremos um exemplo de teste unicaudal a respeito de uma média populacional para o caso em que σ é desconhecido. Uma revista de viagens de negócios quer classificar os aeroportos internacionais de acordo com a avaliação média da população de pessoas que viajam a negócios. Será usada uma escala de classificação, sendo 0 uma avaliação baixa e 10 uma avaliação elevada, e os aeroportos que receberem uma avaliação média populacional maior que 7 serão designados como aeroportos com um atendimento de alto nível.

A equipe da revista pesquisou uma amostra de 60 viajantes de negócios em cada aeroporto para obter os dados da avaliação. A amostra do Aeroporto Heathrow, de Londres, produziu uma avaliação média amostral $\bar{x} = 7{,}25$ e um desvio padrão s da amostra igual a 1,052. Os dados indicam que o Aeroporto Heathrow deveria ser designado como um aeroporto com atendimento de alto nível?

Queremos desenvolver um teste de hipóteses referente a qual decisão de rejeitar H_0 acarretará a conclusão de que a avaliação média populacional do Aeroporto Heathrow seja *maior* que 7. Desse modo, um teste da cauda superior, com H_a: $\mu > 7$, é necessário. As hipóteses nula e alternativa para esse teste da cauda superior são as seguintes:

$$H_0: \mu \leq 7$$
$$H_a \mu > 7$$

Utilizaremos $\alpha = 0{,}05$ como nível de significância para o teste.

Usando a Equação 9.4, com $\bar{x} = 7{,}25$, $s = 1{,}052$ e $n = 60$, o valor da estatística de teste é

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} = \frac{7{,}25 - 7}{1{,}052/\sqrt{60}} = 1{,}84$$

A distribuição amostral de t tem $n - 1 = 60 - 1 = 59$ graus de liberdade. Uma vez que o teste é um teste da cauda superior, o valor p é a área sob a curva da distribuição t à direita de $t = 1{,}84$.

As tabelas de distribuição t apresentadas na maioria dos livros didáticos não conterão detalhes suficientes para determinarmos o valor p exato, como o valor p correspondente a $t = 1{,}84$. Por exemplo, ao usarmos a Tabela 2 do Apêndice B, a distribuição t com 59 graus de liberdade fornece a seguinte informação:

Área da Cauda Superior	0,20	0,10	0,05	0,025	0,01	0,005
Valor t (59 graus de lib.)	0,848	1,296	1,671	2,001	2,391	2,662

Os Apêndices 9.1 e 9.2 explicam como obter o valor p exato com o Minitab ou o Excel.

Vemos que $t = 1{,}84$ está entre 1,671 e 2,001. Não obstante a tabela não fornecer o valor p exato, os valores apresentados na linha "Área da Cauda Superior" indicam que o valor p deve ser menor que 0,05 e maior que 0,025. Com um nível de significância $\alpha = 0{,}05$, essa localização é tudo o que precisamos para saber tomar a decisão de rejeitar a hipótese nula e concluir que o Aeroporto Heathrow deve ser classificado como um aeroporto com atendimento de alto nível.

Softwares como o Minitab e o Excel podem determinar facilmente o valor p exato associado à estatística de teste $t = 1{,}84$. Por exemplo, a saída de dados (*output*) do Minitab da Figura 9.8 apresenta a média amostral $\bar{x} = 7{,}25$, o desvio padrão amostral $s = 1{,}052$ (arredondado), a estatística de teste $t = 1{,}84$ e o valor p exato $= 0{,}035$ referente ao teste de hipóteses da avaliação do Aeroporto Heathrow. Um valor $p = 0{,}035 < 0{,}05$ leva à rejeição da hipótese nula e à conclusão de que o Aeroporto Heathrow deve ser classificado como um aeroporto com atendimento de alto nível. O procedimento passo a passo usado para obtermos a saída do Minitab apresentada na Figura 9.8 é descrito no Apêndice 9.1.

O critério do valor crítico também pode ser usado para se tomar a decisão de rejeição. Com $\alpha = 0{,}05$ e a distribuição t com 59 graus de liberdade, $t_{0,05} = 1{,}671$ é o valor crítico do teste. A regra de rejeição é, portanto,

$$\text{Rejeitar } H_0 \text{ se } t \geq 1{,}671$$

Figura 9.8 Saída de dados do Minitab relativa ao teste de hipóteses da avaliação do Aeroporto Heathrow

```
Test of mu = 7 vs > 7

                                                      95%
                                                     Lower
Variable   N    Mean     StDev    SE Mean    Bound      T       P
Rating    60   7.250   1.05163    0.13577  7.02312   1.84   0.035
```

Com a estatística de teste $t = 1{,}84 \geq 1{,}671$, H_0 é rejeitada, e podemos concluir que o Aeroporto Heathrow pode ser classificado como um aeroporto com atendimento de alto nível.

Teste Bicaudal

Para ilustrar como se realiza um teste bicaudal a respeito de uma média populacional para o caso em que σ é desconhecido, consideremos a situação de teste de hipóteses enfrentada pela Holiday Toys. A empresa manufatura seus produtos e os distribui para mais de mil pontos de revenda. Ao planejar os níveis de produção para a próxima estação de inverno, a Holiday precisa decidir quantas unidades de cada produto deve produzir antes de conhecer a demanda real ao nível de varejo. Em relação ao novo brinquedo mais importante deste ano, o diretor de marketing da Holiday espera que a demanda atinja uma média de 40 unidades por ponto de revenda. Antes de tomar a decisão final de produção baseando-se nessa estimativa, a Holiday decidiu pesquisar uma amostra de 25 varejistas a fim de desenvolver mais informações sobre a demanda pelo novo produto. Cada varejista recebeu informações sobre as características do novo brinquedo, além do custo e do preço de venda sugerido. Depois, cada varejista foi solicitado a especificar um lote de compra previsto.

Considerando que μ denota a média dos lotes de compra da população por ponto de revenda, os dados amostrais serão usados para realizar o seguinte teste de hipótese bicaudal:

$$H_0: \mu = 40$$
$$H_a: \mu \neq 40$$

Se H_0 não puder ser rejeitada, a Holiday continuará seu planejamento da produção baseando-se na estimativa feita pelo diretor de marketing, segundo a qual a média dos lotes de compra da população por ponto de revenda será $\mu = 40$ unidades. Entretanto, se H_0 for rejeitada, a Holiday reavaliará imediatamente seu plano de produção do produto. Um teste de hipóteses bicaudal é usado porque a Holiday quer reavaliar o plano de produção se a média dos lotes de compra da população por ponto de revenda for menor ou maior que o previsto. Uma vez que não há dados históricos disponíveis (trata-se de um novo produto), a média μ da população e o desvio padrão da população devem ser, ambos, estimados usando-se $\bar{x}$ e σ dos dados amostrais.

A amostra de 25 varejistas produziu uma média $\bar{x} = 37,4$ e um desvio padrão $\sigma = 11,79$ unidades. Antes de seguir em frente utilizando a distribuição t, o analista construiu um histograma dos dados amostrais a fim de verificar a forma da distribuição populacional. O histograma dos dados amostrais não apresentou nenhuma evidência de assimetria nem pontos fora da curva extremos, de forma que o analista concluiu que o uso da distribuição t com $n - 1 = 24$ graus de liberdade era apropriado. Usando a Equação 9.4, com $\bar{x} = 37,4$, $\mu_0 = 40$, $s = 11,79$ e $n = 25$, o valor da estatística de teste é:

ARQUIVO DA INTERNET
Orders

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} = \frac{37,4 - 40}{11,79/\sqrt{25}} = -1,10$$

Já que se trata de um teste bicaudal, o valor p é duas vezes a área sob a curva da distribuição t à esquerda de $t = -1,10$. Ao usarmos a Tabela 2 do Apêndice B, notamos que a tabela da distribuição t correspondente a 24 graus de liberdade fornece a seguinte informação:

Área da Cauda Superior	0,20	0,10	0,05	0,025	0,01	0,005
Valor t (24 graus de lib.)	0,857	1,318	1,711	2,064	2,492	2,797

t =/1,10

A tabela de distribuição t contém somente valores t positivos. Entretanto, desde que a distribuição t seja simétrica, podemos encontrar a área sob a curva à direita de $t = 1,10$ e duplicá-la para encontrarmos o valor p. Notamos que $t = 1,10$ está entre 0,857 e 1,318. Na linha "Área da Cauda Superior", notamos que a área na cauda à direita de $t = 1,10$ está entre 0,20 e 0,10. Duplicando esses valores, notamos que o valor p deve estar entre 0,40 e 0,20. Com um nível de significância a = 0,05, agora sabemos que o valor p é maior que α. Portanto, H_0 não pode ser rejeitada. Não há suficientes evidências disponíveis para concluirmos que a Holiday deve alterar seu plano de produção para a próxima estação. Usando o Minitab e o Excel, descobrimos que o valor p exato é 0,282. A Figura 9.9 apresenta as duas áreas sob a curva da distribuição t que fornecem o valor p exato.

A estatística de teste também pode ser comparada com o valor crítico para se tomar a decisão em testes de hipóteses bicaudais. Com $\alpha = 0,05$ e a distribuição t com 24 graus de liberdade, $-t_{0,025} = -2,064$ e $t_{0,025} = 2,064$ são os valores críticos para o teste bicaudal. A regra de rejeição, usando-se a estatística de teste, é:

$$\text{Rejeitar } H_0 \text{ se } t \leq -2,064 \text{ ou se } t \geq 2,064$$

Com base na estatística de teste $t = -1,10$, H_0 não pode ser rejeitada. Esse resultado indica que a Holiday deve manter seu planejamento de produção para a próxima estação baseando-se na expectativa de que $\mu = 40$.

Resumo e Conselho Prático

A Tabela 9.3 apresenta um resumo dos procedimentos de teste de hipóteses a respeito de uma média populacional para o caso em que σ é desconhecido. A diferença fundamental entre esses procedimentos e aqueles em que σ é conhecido é que s é usado em vez de σ no cálculo da estatística de teste. Por esse motivo, a estatística de teste segue a distribuição t.

Figura 9.9 A área sob a curva em ambas as caudas fornece o valor p

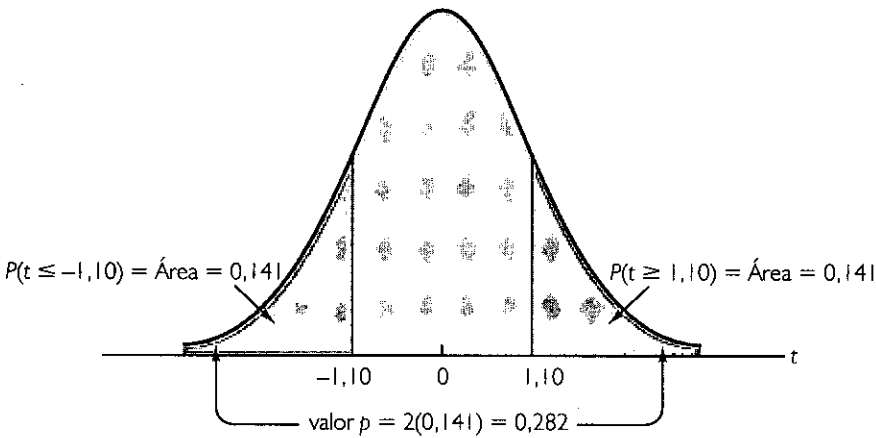



Tabela 9.3 Resumo dos testes de hipótese a respeito de uma média populacional: caso em que σ é desconhecido

	Teste da Cauda Inferior	Teste da Cauda Superior	Teste Bicaudal
Hipótese	H_0: $\mu \geq \mu_0$ H_a: $\mu < \mu_0$	H_0: $\mu \leq \mu_0$ H_a: $\mu > \mu_0$	H_0: $\mu = \mu_0$ H_a: $\mu \neq \mu_0$
Estatística de Teste	$t = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$	$t = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$	$t = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$
Regra de Rejeição: Critério do Valor p	Rejeitar H_0 se o valor $p \leq \alpha$	Rejeitar H_0 se o valor $p \leq \alpha$	Rejeitar H_0 se o valor $p \leq \alpha$
Regra de Rejeição: Critério do Valor Crítico	Rejeitar H_0 se $t \leq -t_\alpha$	Rejeitar H_0 se $t \geq t_\alpha$	Rejeitar H_0 se $t \leq -t_{\alpha/2}$ ou se $t \geq t_{\alpha/2}$

A aplicabilidade dos procedimentos de teste de hipóteses apresentados nesta seção depende da distribuição da população da qual se extrai a amostra e do tamanho da amostra. Quando a população estiver normalmente distribuída, os testes de hipóteses descritos nesta seção produzirão resultados exatos para qualquer tamanho de amostra. Quando a população não estiver normalmente distribuída, os procedimentos serão aproximações. Todavia, observamos que tamanhos de amostras maiores que 50 produzirão bons resultados em quase todos os casos. Se a população for aproximadamente normal, tamanhos de amostra pequenos (por exemplo, $n \leq 15$) podem produzir resultados aceitáveis. Em situações nas quais a população não pode ser aproximada a uma distribuição normal, tamanhos de amostra $n \geq 15$ produzirão resultados aceitáveis contanto que a população não tenha uma assimetria elevada e não contenha pontos fora da curva. Se a população tiver uma assimetria elevada ou se contiver pontos fora da curva, tamanhos de amostra próximos de 50 serão uma boa idéia.

Exercícios

Métodos

23. Considere o seguinte teste de hipótese:

$$H_0: \mu \leq 12$$
$$H_a: \mu > 12$$

Uma amostra de tamanho 25 produziu a média amostral $\bar{x} = 14$ e um desvio padrão amostral $s = 4,32$.

a. Calcule o valor da estatística de teste.

b. O que a tabela de distribuição t (Tabela 2 do Apêndice B) lhe diz sobre o valor p?

c. Com α = 0,05, qual é a sua conclusão?

d. Qual é a regra de rejeição, usando-se o valor crítico? Qual é a sua conclusão?

24. Considere o seguinte teste de hipóteses:

AUTOTESTE

$$H_0: \mu = 18$$
$$H_a: \mu \neq 18$$

Uma amostra de tamanho 48 produziu uma média amostral $\bar{x}$ = 17 e um desvio padrão amostral s = 4,5.

a. Calcule o valor da estatística de teste.

b. O que a tabela de distribuição t (Tabela 2 do Apêndice B) lhe diz sobre o valor p?

c. Com α = 0,05, qual é a sua conclusão?

d. Qual é a regra de rejeição, usando-se o valor crítico? Qual é a sua conclusão?

25. Considere o seguinte teste de hipóteses:

$$H_0: \mu \geq 45$$
$$H_a: \mu < 45$$

Uma amostra de tamanho 36 é usada. Identifique o valor p e apresente sua conclusão em relação a cada um dos seguintes resultados de amostra. Use α = 0,01.

a. $\bar{x}$= 44 e s = 5,2

b. $\bar{x}$= 43 e s = 4,6

c. $\bar{x}$= 46 e s = 5,0

26. Considere o seguinte teste de hipóteses:

$$H_0: \mu = 100$$
$$H_a: \mu = 100$$

Uma amostra de tamanho 65 é usada. Identifique o valor p e apresente sua conclusão em relação a cada um dos seguintes resultados de amostra. Use α = 0,05.

a. $\bar{x}$ = 103 e s = 11,5

b. $\bar{x}$= 96,5 e s = 11,0

c. $\bar{x}$= 102 e s = 10,5

Aplicações

AUTOTESTE

27. A Employment and Training Administration divulgou que a média dos benefícios de seguro-desemprego nos Estados Unidos era de US$ 238 por semana (*The World Almanac*, 2003). Um pesquisador da Virgínia previu que dados amostrais comprovariam que a média dos benefícios de seguro-desemprego na Virgínia estava abaixo do nível nacional.

a. Desenvolva hipóteses apropriadas de tal forma que a rejeição de H_0 sustente a argumentação do pesquisador.

b. Em relação a uma amostra de cem indivíduos, a média amostral dos benefícios de seguro-desemprego semanais foi de US$ 231, com um desvio padrão amostral de US$ 80. Qual é o valor p?

c. Com α = 0,05, qual é a sua conclusão?

d. Repita o teste de hipótese anterior usando o critério do valor crítico.

28. A National Association of Professional Baseball Leagues, Inc. divulgou que o público presente nos jogos das 176 equipes de beisebol da *minor league*[2] atingiu níveis sem precedentes durante a temporada de 2001 (*New York Times*, 28 de julho de 2002). Por jogo, a média de público nos jogos de beisebol da *minor league* foi de 3.530. Na metade da temporada de 2002, o presidente da associação solicitou um relatório de presença do público que esperançosamente mostrasse que a média de público em 2002 ultrapassou o nível de 2001.

a. Formule hipóteses que poderiam ser usadas para determinar se a média de público por jogo em 2002 foi maior que o nível do ano anterior.

---

[2] NT: *Minor league* – Clubes de beisebol profissional não-integrantes das *major leagues*, as duas ligas principais de clubes de beisebol profissional nos Estados Unidos: a *National League* e a *American League*.

b. Suponha que uma amostra de 92 jogos de beisebol da *minor league* disputados durante a primeira metade da temporada de 2002 apresente uma média de público de 3.740 pessoas por jogo, com um desvio padrão amostral igual a 810. Qual é o valor p?
c. Com $\alpha = 0{,}01$, qual é a sua conclusão?

29. O custo de um brilhante de um quilate, com brilho VS2 e cor H, da Diamond Source USA, é de US$ 5.600 (diasource.com, março de 2003). Um joalheiro do meio-oeste liga para seus contatos do *diamond district* de Nova York para verificar se o preço médio dos diamantes lá difere dos US$ 5.600.
 a. Formule hipóteses que possam ser usadas para determinar se a média de preços em Nova York difere dos US$ 5.600.
 b. Suponha que uma amostra de 25 contatos de Nova York produza um preço médio amostral de US$ 5.835 e um desvio padrão amostral de US$ 520. Qual é o valor p?
 c. Com $\alpha = 0{,}05$, a hipótese nula pode ser rejeitada? Qual é a sua conclusão?
 d. Repita o teste de hipótese anterior usando o critério do valor crítico.

30. A CNN, da AOL Time Warner Inc., foi durante muito tempo a líder de audiência em jornalismo da televisão a cabo. A Nielsen Media Research indicou que a média de telespectadores da CNN foi de 600 mil pessoas por dia durante 2002 (*The Wall Street Journal*, 10 de março de 2003). Suponha que, para uma amostra de 40 dias, durante o primeiro semestre de 2003, o público médio tenha sido 612 mil telespectadores, com um desvio padrão de 65 mil pessoas.
 a. Quais são as hipóteses se a gerência da CNN quisesse obter informações sobre quaisquer alterações no público telespectador da CNN?
 b. Qual é o valor p?
 c. Escolha seu próprio nível de significância? Qual é a sua conclusão?
 d. Qual recomendação você faria à gerência da CNN nessa aplicação?

31. A Rafaelis Financial Consulting divulgou que a média trimestral das contas de consumo de água nos Estados Unidos é US$ 47,50 (*U.S. News & World Report*, 12 de agosto de 2002). Alguns sistemas de abastecimento de água são operados por empresas públicas, ao passo que outros sistemas de abastecimento de água são operados por empresas particulares. Um economista destacou que privatização não equivale à competição e que os poderes de monopólio concedidos às empresas públicas agora estão sendo transferidos às empresas privadas. A preocupação é que os consumidores acabem por pagar tarifas maiores que a média pela água fornecida pelas empresas privadas. O sistema de abastecimento de água de Atlanta, na Geórgia, é administrado por uma empresa privada. Uma amostra de 64 consumidores de Atlanta exibiu uma média trimestral de US$ 51 quanto às suas contas de consumo de água, com um desvio padrão amostral igual a US$ 12. Com $\alpha = 0{,}05$, a amostra de consumidores de Atlanta sustenta a conclusão de que existem tarifas acima da média com respeito ao sistema privado de abastecimento de água nessa cidade? Qual é a sua conclusão?

32. De acordo com a National Automobile Dealers Association, o preço médio dos carros usados é US$ 10.192. O gerente de uma revendedora de carros usados de Kansas City revisou uma amostra de 50 vendas recentes de carros usados em sua revendedora, tentando determinar se o preço médio populacional dos carros usados vendidos em sua revendedora em particular diferia da média nacional.
 a. Formule as hipóteses que podem ser usadas para determinar se existe uma diferença na média de preços de carros usados na revendedora.
 b. Qual é o valor p com base em um preço médio amostral de US$ 9.750 e em um desvio padrão amostral de US$ 1.400?
 c. Com $\alpha = 0{,}05$, qual é a sua conclusão?

33. O novo ERC *driver*[3] forjado em titânio, da Callway Golf Company, tem sido descrito como "ilegal" porque promete distâncias de arremesso (*driving distances*) que ultrapassam o padrão estabelecido pela USGA. A *Golf Digest* comparou as distâncias de arremesso reais com o ERC *driver* e com um *driver* aprovado pela USGA, obtendo uma média populacional de distância de arremesso de 256,03 m. Com base em nove arremessos para fins de teste, a média de distância obtida pelo ERC *driver* foi de 262,34 m (*Golf Digest*, 12 de maio de 2000). Responda às questões a seguir supondo um desvio padrão amostral de 9,14 m para a distância de arremesso.

---

[3] NT: *Driver* – Taco de golfe com cabo de madeira e pouca inclinação, usado para lançar a bola do *tee* (ponto a partir do qual se bate a primeira tacada em cada buraco). "ERC" são as iniciais do fundador da Callaway Company: Elly Reeves Callaway (Golfe).

a. Formule as hipóteses nula e alternativa que podem ser usadas para determinar se o novo ERC *driver* tem uma média populacional de distância de arremesso maior que 256,03 m.
b. Em média, quantos metros a mais a bola de golfe percorreu com o ERC *driver*?
c. Com $\alpha = 0,05$, qual é a sua conclusão?

34. A Joan's Nursery é especialista em paisagismo personalizado para áreas residenciais. O custo de mão-de-obra estimado associado a uma proposta de paisagismo em particular baseia-se no número de plantações de árvores, arbustos etc. Para fins de estimação do custo, os gerentes utilizam duas horas de mão-de-obra para o plantio de uma árvore de tamanho médio. Os tempos reais de uma amostra de dez plantações durante o mês passado são apresentados a seguir (o tempo está expresso em horas).

1,7	1,5	2,6	2,2	2,4	2,3	2,6	3,0	1,4	2,3

Com um nível de significância de 0,05, teste se a média de tempo de plantio das árvores difere de duas horas.

a. Estabeleça as hipóteses nula e alternativa.
b. Calcule a média da amostra.
c. Calcule o desvio padrão da amostra.
d. Qual é o valor p?
e. Qual é a sua conclusão?

9.5 PROPORÇÃO DA POPULAÇÃO

Nesta seção, mostramos como realizar um teste de hipóteses a respeito de uma proporção populacional p. Usando p_0 para denotar o valor hipotético da proporção populacional, as três formas de teste de hipóteses a respeito de uma proporção populacional são as seguintes:

$$H_0: p \geq p_0 \qquad H_0: p \leq p_0 \qquad H_0: p = p_0$$
$$H_a: p < p_0 \qquad H_a: p > p_0 \qquad H_a: p \neq p_0$$

A primeira forma é chamada teste da cauda inferior, a segunda forma é denominada teste da cauda superior e a terceira forma é designada teste bicaudal.

Os testes de hipótese a respeito de uma proporção populacional baseiam-se na diferença entre a proporção amostral $\bar{p}$ e a proporção populacional p_0 hipotética. Os métodos utilizados para realizar o teste de hipóteses são similares àqueles que são usados para os testes de hipóteses a respeito de uma média populacional. A única diferença é que usamos a proporção amostral e seu erro padrão para calcular a estatística de teste. O critério do valor p ou o critério do valor crítico é então usado para determinar se a hipótese nula deve ser rejeitada.

Consideremos um exemplo que envolve uma situação enfrentada pelo curso de golfe Pine Creek. No decorrer do ano passado, 20% dos jogadores no Pine Creek eram mulheres. Em um esforço para aumentar a proporção de mulheres jogadoras, o Pine Creek implementou uma promoção especial, idealizada para atrair mulheres golfistas. Um mês depois que a promoção foi implementada, o gerente do curso solicitou um estudo estatístico para determinar se a proporção de mulheres golfistas no Pine Creek havia aumentado. Uma vez que o objetivo do estudo é determinar se a proporção de mulheres golfistas aumentou, um teste da cauda superior, com H_a: $p > 0,20$, é apropriado. As hipóteses nula e alternativa do teste de hipótese do Pine Creek são as seguintes:

$$H_0: p \leq 0,20$$
$$H_a: p > 0,20$$

Se H_0 puder ser rejeitada, os resultados do teste darão apoio estatístico à conclusão de que a proporção de mulheres golfistas aumentou e que a promoção foi benéfica. O gerente do curso especificou que um nível de significância $\alpha = 0,05$ deveria ser usado na execução desse teste de hipóteses.

A etapa seguinte do procedimento de teste de hipóteses é selecionar uma amostra e calcular o valor de uma estatística de teste apropriada. Para mostrar como essa etapa é feita, considerando o teste da cauda superior do Pine Creek, iniciamos com uma discussão geral de como é possível calcular o valor da estatística de teste para qualquer forma de estatística de teste de uma proporção populacional. A distribuição amostral de $\bar{p}$, que é o estimador por ponto do parâmetro populacional p, é a base para desenvolvermos a estatística de teste.

Quando a hipótese nula é verdadeira enquanto igualdade, o valor esperado de $\bar{p}$ equivale ao valor hipotético p_0; ou seja, $E(\bar{p}) = p_0$. O erro padrão de $\bar{p}$ é dado por:

$$\sigma_{\bar{p}} = \sqrt{\frac{p_0(1 - p_0)}{n}}$$

No Capítulo 7, dissemos que se $np \geq 5$ e $n(1 - p) \geq 5$, a distribuição amostral de $\bar{p}$ pode ser aproximada a uma distribuição normal.* Sob essas condições, as quais geralmente se aplicam na prática, a quantidade

$$z = \frac{\bar{p} - p_0}{\sigma_{\bar{p}}} \quad \textbf{(9.5)}$$

tem uma distribuição normal padrão de probabilidade. Com $\sigma_{\bar{p}} = \sqrt{p_0(1 - p_0)/n}$, a variável aleatória z normal padrão é a estatística de teste utilizada para se realizar testes de hipótese a respeito de uma proporção populacional.

ESTATÍSTICA DE TESTE PARA TESTES DE HIPÓTESE A RESPEITO DE UMA PROPORÇÃO POPULACIONAL

$$z = \frac{\bar{p} - p_0}{\sqrt{\frac{p_0(1 - p_0)}{n}}} \quad \textbf{(9.6)}$$

Agora, podemos calcular a estatística de teste correspondente ao teste de hipóteses do Pine Creek. Suponha que uma amostra aleatória de 400 jogadores tenha sido selecionada e que 100 desses jogadores eram mulheres. A proporção de mulheres golfistas é:

$$\bar{p} = \frac{100}{400} = 0{,}25$$

Usando a Equação (9.6), o valor da estatística de teste é:

$$z = \frac{\bar{p} - p_0}{\sqrt{\frac{p_0(1 - p_0)}{n}}} = \frac{0{,}25 - 0{,}20}{\sqrt{\frac{0{,}20(1 - 0{,}20)}{400}}} = \frac{0{,}05}{0{,}02} = 2{,}50$$

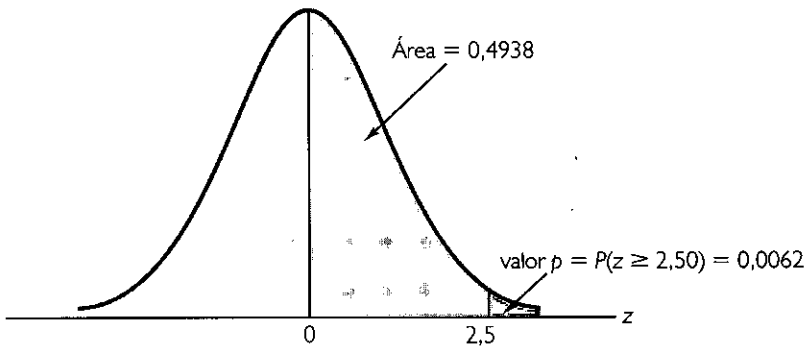
Uma vez que o teste de hipóteses do Pine Creek é um teste da cauda superior, o valor p é a probabilidade de z ser maior ou igual a $z = 2{,}50$; ou seja, é a área sob a curva normal padrão à direita de $z = 2{,}50$. Usando a tabela de áreas da distribuição normal padrão, descobrimos que a área entre a média e $z = 2{,}50$ é 0,4938. Desse modo, o valor p para o teste do Pine Creek é 0,5000 – 0,4938 = 0,0062. A Figura 9.10 apresenta esse cálculo do valor p.

Lembre-se de que o gerente do curso especificou um nível de significância $\alpha = 0{,}05$. Um valor $p = 0{,}0062 < 0{,}05$ fornece suficiente evidência estatística para rejeitarmos H_0 ao nível de significância 0,05. Assim, o teste constitui o suporte estatístico para a conclusão de que a promoção especial aumentou número de jogadoras no curso de golfe Pine Creek.

A decisão de rejeitar ou não rejeitar a hipótese nula também pode ser tomada usando-se o critério do valor crítico. O valor crítico correspondente a uma área de 0,05 na cauda superior de uma distribuição normal padrão é $z_{0,05} = 1{,}645$.

---

* Na maioria das aplicações que envolvem testes de hipótese de uma proporção populacional, os tamanhos de amostra são suficientemente grandes para se usar a aproximação normal. A distribuição amostral exata de $\bar{p}$ é discreta em relação à probabilidade de cada valor de $\bar{p}$ dado pela distribuição binomial. Assim, o teste de hipóteses é um pouco mais complicado para amostras pequenas quando a aproximação normal não pode ser usada.

Figura 9.10 Cálculo do valor p para o teste de hipóteses do Pine Creek

Desse modo, a regra de rejeição usando-se o critério do valor crítico é rejeitar H_0 se $z \geq 1{,}645$. Uma vez que $z = 2{,}50 > 1{,}645$, H_0 é rejeitada.

Novamente, notamos que o critério do valor p e o critério do valor crítico levam à mesma conclusão do teste de hipóteses, mas o critério do valor p fornece mais informação. Com um valor $p = 0{,}0062$, a hipótese nula seria rejeitada para qualquer nível de significância maior ou igual a 0,0062.

Resumo

O procedimento utilizado para realizar um teste de hipóteses a respeito de uma proporção populacional é idêntico ao procedimento utilizado para realizar um teste de hipóteses de uma média populacional. Não obstante somente termos ilustrado a maneira de realizar um teste de hipóteses a respeito de uma proporção populacional para um teste da cauda superior, procedimentos idênticos podem ser usados para testes da cauda inferior e para testes bicaudais. A Tabela 9.4 apresenta um resumo dos testes de hipóteses a respeito de uma proporção populacional.

Tabela 9.4 Resumo dos testes de hipóteses a respeito de uma proporção populacional

	Teste da Cauda Inferior	Teste da Cauda Superior	Teste Bicaudal
Hipóteses	H_0: $\mu \geq \mu_0$ H_a: $\mu < \mu_0$	H_0: $\mu \leq \mu_0$ H_a: $\mu > \mu_0$	H_0: $\mu = \mu_0$ H_a: $\mu \neq \mu_0$
Estatística de Teste	$z = \dfrac{\bar{p} - p_0}{\sqrt{\dfrac{p_0(1-p_0)}{n}}}$	$z = \dfrac{\bar{p} - p_0}{\sqrt{\dfrac{p_0(1-p_0)}{n}}}$	$z = \dfrac{\bar{p} - p_0}{\sqrt{\dfrac{p_0(1-p_0)}{n}}}$
Regra de Rejeição: Critério do Valor p	Rejeitar H_0 se o valor $p \leq \alpha$	Rejeitar H_0 se o valor $p \leq \alpha$	Rejeitar H_0 se o valor $p \leq \alpha$
Regra de Rejeição: Critério do Valor Crítico	Rejeitar H_0 se $z \leq -z_\alpha$	Rejeitar H_0 se $z \geq z_\alpha$	Rejeitar H_0 se $z \leq -z_{\alpha/2}$ ou se $z \geq z_{\alpha/2}$

Exercícios

Métodos

35. Considere o seguinte teste de hipóteses:

$$H_0: p = 0{,}20$$
$$H_a: p \neq 0{,}20$$

Uma amostra de tamanho 400 produziu a proporção amostral $\bar{p} = 0{,}175$.

a. Calcule o valor da estatística de teste.

b. Qual é o valor p?

c. Com $\alpha = 0{,}05$, qual é a sua conclusão?

d. Qual é a regra de rejeição, usando-se o valor crítico? Qual é a sua conclusão?

AUTOTESTE

36. Considere o seguinte teste de hipóteses:

$$H_0: p \geq 0{,}75$$
$$H_a: p < 0{,}75$$

Uma amostra de 300 itens foi selecionada. Calcule o valor p e apresente a sua conclusão com respeito a cada um dos seguintes resultados amostrais. Use $\alpha = 0{,}05$.

a. $\bar{p} = 0{,}68$

b. $\bar{p} = 0{,}72$

c. $\bar{p} = 0{,}70$

d. $\bar{p} = 0{,}77$

Aplicações

37. O Heldrich Center for Workforce Development revelou que 40% dos usuários de internet recebiam mais de dez mensagens de e-mail por dia (*USA Today*, 7 de maio de 2000). Um estudo similar sobre o uso de e-mails foi repetido em 2002.

a. Formule as hipóteses que podem ser usadas para determinar se a proporção de usuários de internet que recebem mais de dez mensagens de e-mail por dia aumentou.

b. Se uma amostra de 425 usuários de internet revelou que 189 pessoas recebem mais de dez mensagens de e-mail por dia, qual é o valor p?

c. Com $\alpha = 0{,}05$, qual é a sua conclusão?

AUTOTESTE

38. Um estudo realizado pela *Consumer Reports* mostrou que 64% das pessoas que fazem compras em supermercados acreditam que as marcas dos próprios supermercados são tão boas quanto as marcas de renome nacional. Para investigar se esse resultado se aplica ao seu próprio produto, o fabricante de uma marca de *ketchup* reconhecida nacionalmente perguntou a uma amostra de compradores se eles acreditavam que o *ketchup* de supermercado era tão bom quanto aquele de renome nacional.

a. Formule as hipóteses que poderiam ser usadas para determinar se a porcentagem de pessoas que fazem compras em supermercados e que acreditam que o *ketchup* de supermercado era tão bom quanto o *ketchup* de marca nacional diferia de 64%.

b. Se uma amostra de 100 compradores revelasse 52 pessoas que declaram que a marca de supermercado era tão boa quanto a marca nacional, qual é o valor p?

c. Com $\alpha = 0{,}05$, qual é a sua conclusão?

d. O fabricante de *ketchup* de marca nacional deve ficar satisfeito com essa conclusão? Explique.

39. O National Center for Health Statistics publicou um relatório que afirmava que 70% dos adultos não se exercitam regularmente (Associated Press, 7 de abril de 2002). Um pesquisador decidiu realizar um estudo para verificar se a afirmação do National Center for Health Statistics diferia em termos de estado para estado.

a. Estabeleça as hipóteses nula e alternativa supondo que a intenção do pesquisador seja de identificar os estados que diferem dos 70% relatados pelo National Center for Health Statistics.

b. Com $\alpha = 0{,}05$, qual é a conclusão da pesquisa para os seguintes estados:

Wisconsin: 252 de 350 adultos não se exercitavam regularmente
Califórnia: 189 de 300 adultos não se exercitavam regularmente

40. Antes do Super Bowl[4] de 2003, a rede ABC previu que 22% do público do Super Bowl manifestaria interesse em assistir a um dos seus novos programas de televisão a serem exibidos em breve, incluindo "8 Simple Rules", "Are You Hot?" e "Dragnet". A ABC exibiu comerciais desses novos programas de televisão durante o Super Bowl. No dia seguinte ao Super Bowl, o Intermediate Advertising Group, de Nova York, tomou uma amostra de 1.532 telespectadores que viram os comerciais e revelou que 414 disseram que assistiriam a um dos anunciados programas de televisão da ABC (*The Wall Street Journal*, 30 de janeiro de 2003).

[4] NT: Super Bowl (*ou* Superbowl) – Final do campeonato de futebol norte-americano.

a. Qual é a estimação por ponto da proporção do público que disse que assistiria aos programas de televisão depois de verem os comerciais de TV?
b. Com $\alpha = 0{,}05$, determine se a intenção de assistir aos programas de televisão da ABC se elevou significativamente depois de verem os comerciais de televisão. Formule as hipóteses apropriadas, calcule o valor p e apresente a sua conclusão.
c. Por que esses estudos são valiosos para as empresas e para as firmas de publicidade?

41. O Microsoft Outlook é o gerenciador de e-mails mais amplamente usado. Um executivo da Microsoft afirma que o Microsoft Outlook é utilizado por, no mínimo, 75% dos usuários de internet. Uma amostra de usuários de internet será usada para testar essa afirmação.
a. Formule as hipóteses que podem ser usadas para testar a afirmação.
b. Um estudo realizado pela Merryll Lynch relatou que o Microsoft Outlook é usado por 72% dos usuários de internet (CNBC, junho de 2000). Suponha que o relatório tenha se baseado em um tamanho de amostra de 300 usuários de internet. Qual é o valor p?
c. Com $\alpha = 0{,}05$, a afirmação do executivo referente a, "no mínimo, 75%" deve ser rejeitada?

42. De acordo com a American Housing Survey, do Departamento do Censo dos Estados Unidos, a razão principal que leva as pessoas que mudam de residência a escolherem determinada região é o fato de a localização ser conveniente para o trabalho (*USA Today*, 24 de dezembro de 2002). Com base nos dados do Departamento do Censo de 1990, sabemos que 24% das pessoas que mudaram de residência indicaram "localização conveniente para o trabalho" como a razão principal para escolherem a nova região. Suponha que uma amostra de 300 pessoas que se mudaram durante 2003 tenha revelado que 93 o fizeram com o objetivo de morar mais perto do trabalho. Os dados da amostra dão suporte à conclusão de pesquisa segundo a qual em 2003 um número maior de pessoas escolheu onde morar baseando-se em quão perto estarão do trabalho? Qual é a estimação por ponto da proporção de pessoas que se mudaram em 2003 que escolheram a nova região porque a localização é conveniente para o trabalho? Qual é a sua conclusão de pesquisa? Use $\alpha = 0{,}05$.

43. Um artigo sobre a maneira de dirigir publicado no município de Strathcona, em Alberta, no Canadá, afirmou que 48% dos motoristas não paravam nos cruzamentos com sinal fechado nas estradas do município (*Edmonton Journal*, 19 de julho de 2000). Dois meses mais tarde, um estudo de acompanhamento coletou dados a fim de verificar se essa porcentagem se modificara.
a. Formule as hipóteses para determinar se a proporção dos motoristas que não paravam nos cruzamentos com sinal fechado havia modificado.
b. Suponha que o estudo tenha revelado que 360 dentre 800 motoristas não paravam nos cruzamentos com sinal fechado. Qual é a proporção amostral Qual é o valor p?
c. Com $\alpha = 0{,}05$, qual é a sua conclusão?

44. Em uma matéria de capa, a *Business Week* publicou informações a respeito dos hábitos de dormir dos norte-americanos (*Business Week*, 26 de janeiro de 2004). O artigo afirmou que a privação do sono leva a uma série de problemas e apontou que o deixar de dormir provoca acidentes fatais nas estradas. Cinqüenta e um por cento dos motoristas adultos admitem dirigir enquanto estão sonolentos. Um pesquisador aventou a hipótese de que essa questão era um problema ainda maior para as pessoas que trabalham em turnos da noite.
a. Formule as hipóteses que podem ser usadas para ajudar a determinar se mais de 51% da população de trabalhadores do turno da noite admitem dirigir enquanto estão sonolentos.
b. Uma amostra de 500 trabalhadores do turno da noite revelou que 232 admitiram dirigir enquanto estavam sonolentos. Qual é a proporção amostral? Qual é o valor p?
c. Com $\alpha = 0{,}01$, qual é a sua conclusão?

45. A Drugstore.com foi a primeira empresa de comércio eletrônico a oferecer produtos de farmácia e perfumaria a varejo pela internet. Os clientes da Drugstore.com tinham a oportunidade de comprar produtos para a saúde, beleza, cuidados pessoais, bem-estar e farmacêuticos pela internet. Ao final de dez meses de operação a empresa relatou que 44% das encomendas eram feitas por clientes que já haviam comprado anteriormente (*Drugstore.com Annual Report*, 2 de janeiro de 2000). Suponha que a Drugstore.com use uma amostra de encomendas de clientes a cada trimestre para determinar se a proporção de encomendas de clientes que já compraram anteriormente se modificou do $p = 0{,}44$ original.

a. Formule as hipóteses nula e alternativa.
b. Durante o primeiro trimestre, uma amostra de 500 encomendas exibiu 205 clientes que já haviam comprado anteriormente. Qual é o valor p? Use $\alpha = 0{,}05$. Qual é a sua conclusão?
c. Durante o segundo trimestre, uma amostra de 500 encomendas exibiu 245 clientes que já haviam comprado anteriormente. Qual é o valor p? Use $\alpha = 0{,}05$. Qual é a sua conclusão?

Resumo

O teste de hipóteses é um procedimento estatístico que usa dados amostrais para determinar se a afirmação a respeito do valor de um parâmetro populacional deve ou não ser rejeitada. As hipóteses são duas afirmações antagônicas sobre um parâmetro populacional. Uma afirmação se denomina hipótese nula (H_0), e a outra, hipótese alternativa (H_a). Na Seção 9.1, apresentamos diretrizes para o desenvolvimento de hipóteses relativas a três situações que são encontradas freqüentemente na prática.

Quando se quiser que dados históricos ou outras informações constituam uma base para se supor que o desvio padrão da população seja conhecido, o procedimento de teste de hipóteses se baseará na distribuição normal padrão. Quando se quiser que σ seja desconhecido, o desvio padrão s da amostra será usado para estimar σ e o procedimento de teste de hipóteses se baseará na distribuição t. Em ambos os casos, a qualidade dos resultados depende tanto da forma da distribuição populacional quanto do tamanho da amostra. Se a população tiver uma distribuição normal, ambos os procedimentos de teste de hipóteses serão aplicáveis, até mesmo com tamanhos de amostra pequenos. Se a população não estiver normalmente distribuída, tamanhos de amostra maiores serão necessários. Diretrizes gerais sobre o tamanho da amostra foram apresentadas nas Seção 9.3 e 9.4. No caso de testes de hipóteses a respeito de uma proporção populacional, o procedimento de testes de hipóteses utiliza uma estatística de teste baseada na distribuição normal padrão.

Em todos os casos, o valor da estatística de teste é usado para calcular um valor p para o teste. O valor p é uma probabilidade, calculada usando-se a estatística de teste, que mede o suporte (ou a falta de suporte) que a amostra dá à hipótese nula. Se o valor p for menor ou igual ao nível de significância α, a hipótese nula poderá ser rejeitada.

As conclusões do teste de hipóteses também podem ser tomadas ao comparar-se o valor da estatística de teste com um valor crítico. Quanto aos testes da cauda inferior, a hipótese nula é rejeitada se o valor da estatística de teste for menor ou igual ao valor crítico. Em relação aos testes da cauda superior, a hipótese nula é rejeitada se o valor da estatística de teste for maior ou igual ao valor crítico. Os testes bicaudais consistem em dois valores críticos: uma na cauda inferior da distribuição amostral e um na cauda superior. Nesse caso, a hipótese nula é rejeitada se o valor da estatística de teste for menor ou igual ao valor crítico na cauda inferior ou maior ou igual ao valor crítico na cauda superior.

Glossário

Hipótese nula A hipótese experimentalmente considerada verdadeira no procedimento de teste de hipóteses.

Hipótese alternativa A hipótese considerada verdadeira se a hipótese nula for rejeitada.

Erro do Tipo I O erro de rejeitar H_0 quando ela é verdadeira.

Erro do Tipo II O erro de aceitar H_0 quando ela é falsa.

Nível de significância A probabilidade de se cometer um erro do Tipo I quando a hipótese nula é verdadeira enquanto igualdade.

Teste unicaudal Um teste de hipóteses no qual a rejeição da hipótese nula ocorre para valores da estatística de teste em uma cauda de sua distribuição amostral.

Estatística de teste Uma estatística cujo valor ajuda a determinar se a hipótese nula pode ser rejeitada.

Valor p Uma probabilidade, calculada usando-se a estatística de teste, que mede o suporte (ou a falta de suporte) que a amostra dá à hipótese nula. Quanto a um teste da cauda inferior, o valor p é a probabilidade de se obter um valor para a estatística de teste tão pequeno ou menor que aquele que é fornecido pela amostra. Em relação a um teste da cauda superior, o valor p é a probabilidade de se obter um valor para a estatística de teste tão grande ou maior que aquele que é fornecido pela amostra. Para um teste bicaudal, o valor p é a probabilidade de se obter um valor para a estatística de teste tão improvável ou mais improvável que aquele que é fornecido pela amostra.

Valor crítico Um valor que é comparado com a estatística de teste para determinar se H_0 deve ser rejeitada.

Teste bicaudal Um teste de hipóteses no qual a rejeição da hipótese nula ocorre para valores da estatística de teste em qualquer uma das caudas de sua distribuição amostral.

Fórmulas-Chave

Estatística de Teste para Testes de Hipóteses a Respeito de uma Média Populacional: σ Conhecido

$$z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} \tag{9.1}$$

Estatística de Teste para Testes de Hipóteses a Respeito de uma Média Populacional: σ Desconhecido

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} \tag{9.4}$$

Estatística de Teste para Testes de Hipóteses a Respeito de uma Proporção Populacional

$$z = \frac{\bar{p} - p_0}{\sqrt{\frac{p_0(1 - p_0)}{n}}} \tag{9.6}$$

Exercícios Suplementares

46. Uma linha de produção opera com um peso médio de preenchimento de 453 g por recipiente. Preencher com um volume maior ou com um volume menor constitui um sério problema e, quando é detectado, exige que o operador interrompa a linha de produção para reajustar o mecanismo de enchimento. A partir de dados passados, presume-se um desvio padrão populacional $\sigma = 22{,}67$ g. Um inspetor de controle da qualidade seleciona uma amostra de 30 itens a cada hora e, nesse momento, toma a decisão de interromper ou não a linha de produção para fazer o reajuste. O nível de significância é $\alpha = 0{,}05$.
   a. Estabeleça o teste de hipótese para essa aplicação de controle da qualidade.
   b. Se uma média amostral $\bar{x} = 462{,}6$ g tiver sido encontrada, qual é o valor p? Qual ação você recomendaria?
   c. Se uma média amostral $\bar{x} = 448{,}4$ g tiver sido encontrada, qual é o valor p? Qual ação você recomendaria?
   d. Use o critério do valor crítico. Qual é a regra de rejeição para o procedimento de teste de hipóteses apresentado anteriormente? Repita os itens (b) e (c). Você chegaria à mesma conclusão?

47. Na Western University, a média histórica das pontuações nos exames para obtenção de bolsas de estudo correspondente às inscrições feitas por calouros é 900. Presume-se que o desvio padrão histórico da população $\sigma = 180$ seja conhecido. Anualmente, o vice-reitor usa uma amostra das inscrições para determinar se a média de pontuação nos exames correspondente às inscrições dos calouros se modificou.
   a. Estabeleça as hipóteses.
   b. Qual é a estimação por intervalo de confiança de 95% da média populacional de pontuação nos exames se uma amostra de 200 inscrições tiver produzido uma média amostral $\bar{x} = 935$?
   c. Use o intervalo de confiança para realizar um teste de hipóteses. Usando $\alpha = 0{,}05$, qual é a sua conclusão?
   d. Qual é o valor p?

48. O salário anual médio da população de professores do ensino público no Estado de Nova York é US$ 45.250. Uma média amostral do salário anual médio da população de professores do ensino público na cidade de Nova York é US$ 47 mil (*Time*, 3 de abril de 2000). Suponha que os resultados relativos à cidade de Nova York se baseiem em uma amostra de 95 professores. Suponha que o desvio padrão σ da população seja US$ 6.300.
   a. Formule as hipóteses nula e alternativa que podem ser usadas para determinar se os dados amostrais sustentam a conclusão de que os professores do ensino público da cidade de Nova York têm uma média salarial mais elevada do que os professores do ensino público do estado de Nova York.
   b. Qual é o valor p?
   c. Use $\alpha = 0{,}01$. Qual é a sua conclusão?

49. De acordo com a National Association of Colleges and Employers, no ano 2000, o salário médio anual dos graduados em contabilidade formados em Administração era de US$ 37 mil (*Time*, 8 de maio de 2000). Em um estudo de acompanhamento realizado em junho de 2001, uma amostra de 48 graduados com *major*[5] em Contabilidade produziu uma média amostral de US$ 38.100 e um desvio padrão de US$ 5.200.

    a. Formule as hipóteses nula e alternativa que podem ser usadas para determinar se os dados amostrais sustentam a conclusão de que os graduados em Contabilidade em 2001 tinham um salário médio maior que o salário médio anual de US$ 37 mil no ano 2000.
    b. Qual é o valor p?
    c. Use $\alpha = 0{,}05$. Qual é a sua conclusão?

50. O College Board divulgou que o número médio de inscrições para o primeiro ano nos colégios e universidades públicos é igual a 6 mil (*USA Today*, 26 de dezembro de 2002). Durante um período de inscrição/matrícula recente, uma amostra de 32 colégios e universidades revelou que o número médio amostral de inscrições para o primeiro ano foi de 5.812, com um desvio padrão amostral de 1.140. Os dados indicam uma alteração no número médio de inscrições? Use $\alpha = 0{,}05$.

51. Um extenso estudo do custo de assistência médica nos Estados Unidos apresentou dados que mostram que a média de gastos por segurado do Medicare em 2003 foi de US$ 6.883 (*Money*, outono de 2003). Para investigar possíveis diferenças no país, um pesquisador tomou uma amostra de 40 segurados do Medicare em Indianápolis. Quanto à amostra de Indianápolis, a média de gastos com o Medicare em 2003 foi de US$ 5.980 e o desvio padrão foi de US$ 2.518.

    a. Estabeleça as hipóteses que seriam usadas se quiséssemos determinar se a média anual de gastos com o Medicare em Indianápolis é menor que a média nacional.
    b. Use os resultados amostrais apresentados anteriormente para calcular a estatística de teste e o valor p.
    c. Use $\alpha = 0{,}05$. Qual é a sua conclusão?
    d. Repita o teste de hipótese usando o critério do valor crítico.

52. A câmara de comércio de uma comunidade litorânea do Golfo da Flórida anuncia que uma propriedade residencial na região está disponível a um custo médio de US$ 125 mil ou menos por lote. Suponha que uma amostra de 32 propriedades forneça uma média amostral de US$ 130 mil por lote e um desvio padrão amostral de US$ 12.500. Usando um nível de significância de 0,05, teste a validade da afirmação feita no anúncio.

53. O rendimento médio por ação da população de corporações de serviços financeiros, incluindo a American Express, o E*TRADE Group, a Goldman Sachs e a Merrill Lynch, foi de US$ 3 (*Business Week*, 14 de agosto de 2000). Em 2001, uma amostra de 10 corporações de serviços financeiros forneceu os seguintes dados de rendimento por ação:

    1,92 2,16 3,63 3,16 4,02 3,14 2,20 2,34 3,05 2,38

    a. Formule as hipóteses nula e alternativa que podem ser usadas para determinar se o rendimento médio por ação da população em 2001 difere dos US$ 3 registrados em 2000.
    b. Calcule a média amostral.
    c. Calcule o desvio padrão da amostra.
    d. Qual é o valor p?
    e. Use $\alpha = 0{,}05$. Qual é a sua conclusão?

54. Um estudo realizado pela organização Centers for Disease Control (CDC) revelou que 23,3% dos adultos são fumantes e que aproximadamente 70% dos que fumam indicam que querem parar de fumar (Associated Press, 26 de julho de 2002). A CDC relatou que, das pessoas que fumaram em algum período da vida, 50% conseguiram abandonar o hábito. Parte do estudo sugeriu que o índice de sucesso para deixar de fumar se elevava de acordo com o nível de educação. Suponha que uma amostra de 100 graduados em cursos superiores que fumaram em algum período da vida tenha revelado que 64 foram capazes de parar de fumar de maneira bem-sucedida.

[5] NT: *Major Educ.* – Designa uma matéria ou área de estudo na qual o estudante se especializa e se gradua (Estados Unidos).

a. Estabeleça as hipóteses que podem ser usadas para determinar se a população de graduados em cursos superiores apresenta um índice maior que a população global quando se trata de vencer o hábito de fumar.
b. Considerando os dados amostrais, qual é a proporção de graduados em cursos superiores que, tendo fumado em algum período da vida, foram capazes de parar de fumar?
c. Qual é o valor p? Com $\alpha = 0{,}01$, qual é a conclusão do seu teste de hipóteses?

55. Uma promoção de uma empresa aérea para pessoas que fazem viagens de negócios baseia-se na suposição de que dois terços dessas pessoas usam computadores *laptop* em viagens de negócios noturnas.
    a. Estabeleça as hipóteses que podem ser usadas para testar a suposição.
    b. Qual é a proporção amostral de uma pesquisa patrocinada pela American Express que revelou que 355 de 546 pessoas que fazem viagens de negócios usam um computador *laptop* em viagens de negócios noturnas?
    c. Qual é o valor p?
    d. Use $\alpha = 0{,}05$. Qual é a sua conclusão?
56. Os funcionários de escritório da Shell Oil foram solicitados a responder qual programação de trabalho seria a mais atraente: trabalhar cinco dias de oito horas por semana ou trabalhar quatro dias de dez horas por semana (*USA Today*, 11 de setembro de 2000). Admitamos que p = a proporção da população de funcionários de escritório que preferem trabalhar quatro dias de dez horas por semana.
    a. Estabeleça as hipóteses para o caso de a gerência da Shell estar interessada em obter evidências estatísticas que mostrem que mais de 50% dos funcionários de escritório preferem trabalhar quatro dias de dez horas por semana.
    b. Qual é a proporção amostral se uma amostra de 105 funcionários de escritório tiver revelado que 67 prefeririam a programação de quatro dias de dez horas?
    c. Qual é o valor p? Use $\alpha = 0{,}01$. Qual é a sua conclusão?
57. Durante o ano eleitoral de 2004, novos resultados de pesquisa de opinião eram publicados diariamente. Em uma pesquisa realizada pela IBD/TIPP com 910 adultos, 503 entrevistados revelaram que estavam otimistas quanto ao cenário nacional, e o índice de popularidade do presidente Bush saltou 4,7 pontos, atingindo 55,3 (*Investor's Business Daily*, 14 de janeiro de 2004).
    a. Qual é a proporção amostral dos entrevistados que estavam otimistas em relação ao cenário nacional?
    b. Um gerente de campanha quer afirmar que essa pesquisa indica que a maioria dos adultos está otimista quanto ao cenário nacional. Construa um teste de hipóteses de forma que a rejeição da hipótese nula possibilite a conclusão de que a proporção otimista é maior que 50%.
    c. Use os dados da pesquisa de opinião para calcular o valor p para o teste de hipóteses do item (b). Explique ao gerente o que significa o valor p em relação ao nível de significância dos resultados.
58. Uma estação de rádio de Myrtle Beach anunciou que pelo menos 90% dos hotéis e motéis estariam lotados no fim de semana do Memorial Day. A estação aconselhou os ouvintes a fazerem reservas antecipadamente, caso planejassem passar o fim de semana no balneário. No sábado à noite, uma amostra de 58 hotéis e motéis revelou que 49 exibiam o anúncio "sem vagas" e 9 "com vagas". Qual é a sua reação à afirmação da estação de rádio depois de ver a evidência da amostra? Use $\alpha = 0{,}05$ ao realizar o teste estatístico. Qual é o valor p?
59. Os indicadores de saúde ambiental incluem a qualidade do ar, a qualidade da água e a qualidade dos alimentos. Há 25 anos, 47% das amostras de alimentos norte-americanas continham resíduos de defensivos agrícolas (*U.S. News & World Report*, 17 de abril de 2000). Em um estudo recente, 44 de 125 amostras de alimentos continham resíduos de defensivos agrícolas.
    a. Estabeleça as hipóteses que podem ser usadas para mostrar que a proporção populacional sofreu um declínio.
    b. Qual é a proporção amostral?
    c. Qual é o valor p?
    d. Use $\alpha = 0{,}01$. Qual é a sua conclusão?

Estudo de Caso 1 – Quality Associates, Inc.

A Quality Associates, Inc., uma firma de consultoria, orienta seus clientes a respeito de procedimentos amostrais e estatísticos que podem ser usados para controlar seus processos de manufatura. Em uma apli-

cação em particular, um cliente deu à Quality Associates uma amostra de 800 observações feitas durante um período em que o processo do cliente estava operando satisfatoriamente. O desvio padrão da amostra para esses dados era 0,21; portanto, com tantos dados, presumiu-se que o desvio padrão da população fosse 0,21. A Quality Associates sugeriu então que amostras aleatórias de tamanho 30 fossem tomadas periodicamente para monitorar o processo em base contínua. Ao analisar as novas amostras, o cliente poderia saber rapidamente se o processo estava operando satisfatoriamente. Quando o processo não estivesse operando satisfatoriamente, medidas corretivas poderiam ser tomadas para eliminar o problema. A especificação do projeto indicava que a média do processo devia ser 12. O teste de hipóteses sugerido pela Quality Associates foi o seguinte:

$$H_0: \mu = 12$$
$$H_a: \mu \neq 12$$

Medidas corretivas serão tomadas sempre que H_0 for rejeitada.

As amostras a seguir foram coletadas em intervalos horários durante o primeiro dia de operação do novo procedimento de controle estatístico do processo. Esses dados estão disponíveis no conjunto de dados (*data set*) Quality.

ARQUIVO DA INTERNET

Quality

Amostra 1	Amostra 2	Amostra 3	Amostra 4
11,55	11,62	11,91	12,02
11,62	11,69	11,36	12,02
11,52	11,59	11,75	12,05
11,75	11,82	11,95	12,18
11,90	11,97	12,14	12,11
11,64	11,71	11,72	12,07
11,80	11,87	11,61	12,05
12,03	12,10	11,85	11,64
11,94	12,01	12,16	12,39
11,92	11,99	11,91	11,65
12,13	12,20	12,12	12,11
12,09	12,16	11,61	11,90
11,93	12,00	12,21	12,22
12,21	12,28	11,56	11,88
12,32	12,39	11,95	12,03
11,93	12,00	12,01	12,35
11,85	11,92	12,06	12,09
11,76	11,83	11,76	11,77
12,16	12,23	11,82	12,20
11,77	11,84	12,12	11,79
12,00	12,07	11,60	12,30
12,04	12,11	11,95	12,27
11,98	12,05	11,96	12,29
12,30	12,37	12,22	12,47
12,18	12,25	11,75	12,03
11,97	12,04	11,96	12,17
12,17	12,24	11,95	11,94
11,85	11,92	11,89	11,97
12,30	12,37	11,88	12,23
12,15	12,22	11,93	12,25

Relatório Administrativo

1. Realize um teste de hipóteses para cada amostra com um nível de significância 0,01 e determine quais medidas, se for o caso, devem ser tomadas. Forneça a estatística de teste e o valor p para cada teste.
2. Calcule o desvio padrão de cada uma das quatro amostras. A suposição de 0,21 para o desvio padrão da população parece razoável?
3. Calcule os limites para a média amostral $\bar{x}$ em torno de $\mu = 12$ de forma que, estando a nova média amostral dentro desses limites, considere que o processo esteja funcionando satisfatoriamente. Se

$\bar{x}$ ultrapassar o limite máximo ou se $\bar{x}$ ficar abaixo do limite mínimo, medidas corretivas serão tomadas. Esses limites se denominam limites máximo e mínimo de controle para fins de controle da qualidade.

4. Discuta as implicações de se mudar o nível de significância para um valor maior. Qual equívoco ou erro poderia se avolumar se o nível de significância fosse aumentado?

Estudo de Caso 2 – Estudo do Desemprego

Mensalmente, o U.S. Bureau of Labor Statistics publica uma série de estatísticas sobre o número de pessoas que estão desempregadas nos Estados Unidos e a média de tempo em que estão desempregadas. Em relação a novembro de 1988, o Bureau of Labor Statistics divulgou que a duração média nacional de desemprego era 14,6 semanas.

O prefeito de Filadélfia solicitou um estudo sobre a situação de desemprego na cidade. Uma amostra de 50 habitantes desempregados de Filadélfia incluiu dados sobre a idade e o número de semanas em que estavam sem emprego.

Apresentamos a seguir uma parte dos dados coletados em novembro de 1998. O conjunto de dados completo está disponível no arquivo BLS.

Idade	Semanas	Idade	Semanas	Idade	Semanas
56	22	22	11	25	12
35	19	48	6	25	1
22	7	48	22	59	33
57	37	25	5	49	26
40	18	40	20	33	13

ARQUIVO DA INTERNET
BLS

Relatório Administrativo

1. Use estatística descritiva para resumir os dados.
2. Desenvolva uma estimação por intervalo de confiança de 95% da média de idade das pessoas desempregadas em Filadélfia.
3. Realize um teste de hipóteses para determinar se a duração média do desemprego em Filadélfia é maior que a duração média nacional de 14,6 semanas. Use um nível de significância de 0,01. Qual é a sua conclusão?
4. Há uma relação entre a idade do indivíduo desempregado e o número de semanas de desemprego? Explique.

Apêndice 9.1 – Testes de Hipóteses com o Minitab

Descrevemos o uso do Minitab para realizar testes de hipótese a respeito de uma média populacional e de uma proporção populacional.

Média da População: σ Conhecido

Ilustramos nossa exposição usando o exemplo da distância percorrida pela bola de golfe MaxFlight apresentado na Seção 9.3. Os dados estão na coluna C1 de uma planilha do Minitab. Consideramos que o desvio padrão populacional $\sigma = 12$ seja conhecido e que o nível de significância seja $\alpha = 0{,}05$. As etapas a seguir podem ser usadas para testar a hipótese H_0: $\mu = 295$ contra H_a: $\mu \neq 295$.

ARQUIVO DA INTERNET
GolfTest

Etapa 1. Selecione o menu **Stat**
Etapa 2. Escolha **Basic Statistics**
Etapa 3. Escolha **1-Sample Z**
Etapa 4. Quando a caixa de diálogo 1-Sample Z aparecer:
Digite C1 na caixa **Samples in columns**
Digite 12 na caixa **Standard deviation**
Digite 295 na caixa **Test mean**
Selecione **Options**

Etapa 5. Quando a caixa de diálogo1-Sample Z-Options aparecer:
Digite 95 na caixa **Confidence level***
Selecione **not equal** na caixa **Alternative**
Dê um clique em **OK**
Etapa 6. Dê um clique em **OK**

Além dos resultados do teste de hipóteses, o Minitab fornece um intervalo de confiança de 95% relativo à média da população.

O procedimento pode ser facilmente modificado para um teste de hipóteses unicaudal ao selecionar-se a opção **less than** ou **greater than** na caixa **Alternative** na etapa 5.

Média da População: σ Desconhecido

ARQUIVO DA INTERNET
AirRating

As avaliações que 60 viajantes de negócios deram ao Aeroporto Heathrow foram inseridas na coluna C1 de uma planilha do Minitab. O nível de significância para o teste é $\alpha = 0{,}05$, e o desvio padrão σ da população será estimado pelo desvio padrão s da amostra. As etapas a seguir podem ser usadas para testar as hipóteses H_0: $\mu \leq 7$ contra H_a: $\mu > 7$.

Etapa 1. Selecione o menu **Stat**
Etapa 2. Escolha **Basic Statistics**
Etapa 3. Escolha **1-Sample t**
Etapa 4. Quando a caixa de diálogo 1-Sample t aparecer:
Digite C1 na caixa **Samples in columns**
Digite 7 na caixa **Test mean**
Selecione **Options**
Etapa 5. Quando a caixa de diálogo 1-Sample t aparecer:
Digite 95 na caixa **Confidence level***
Selecione **greater than** na caixa **Alternative**
Dê um clique em **OK**
Etapa 6. Dê um clique em **OK**

O estudo de avaliação do Aeroporto Heathrow envolveu uma hipótese alternativa "maior que". As etapas anteriores podem ser facilmente modificadas para outros testes de hipótese ao selecionar-se as opções **less than** ou **not equal** na caixa **Alternative** na etapa 5.

Proporção da População

ARQUIVO DA INTERNET
WomenGolf

Ilustramos nossa exposição usando o exemplo do curso de golfe Pine Creek apresentado na Seção 9.5. Os dados com as respostas Female (Mulher) e Male (Homem) estão na coluna C1 de uma planilha do Minitab. O Minitab usa uma classificação em ordem alfabética para as respostas e seleciona a *segunda resposta* da proporção populacional de interesse. Neste exemplo, o Minitab usa a classificação em ordem alfabética Female-Male (Mulher-Homem) para fornecer os resultados correspondentes à proporção populacional de respostas Male (Homem). Uma vez que Female (Mulher) é a resposta de interesse, modificamos a ordem de classificação do Minitab da seguinte maneira: Selecione qualquer célula da coluna e use a seqüência Editor > Column > Value Order. Depois escolha a opção de introduzir uma ordem especificada pelo usuário. Certifique-se de que as respostas estão classificadas na ordem Male-Female (Homem-Mulher) na caixa **Define-an-Order**. A rotina 1 Proportion do Minitab fornecerá então os resultados do teste de hipótese correspondentes à proporção populacional de golfistas mulheres. Prosseguimos da seguinte maneira:

Etapa 1. Selecione o menu **Stat**
Etapa 2. Escolha **Basic Statistics**
Etapa 3. Escolha **1 Proportion**

* O Minitab fornece simultaneamente os resultados do teste de hipóteses e os resultados de estimação por intervalo. O usuário pode selecionar qualquer nível de confiança para a estimação por intervalo da média populacional: aqui, sugerimos um intervalo de confiança de 95%.

Etapa 4. Quando a caixa de diálogo1 Proportion aparecer:
Digite C1 na caixa **Samples in Columns**
Selecione **Options**

Etapa 5. Quando a caixa de diálogo 1 Proportion-Options aparecer:
Digite 95 na caixa **Confidence level***
Selecione 0,20 na caixa **Test proportion**
Selecione **greater than** na caixa **Alternative**
Selecione **Use test and interval based on normal distribution**
Dê um clique em **OK**

Etapa 6. Dê um clique em **OK**

Apêndice 9.2 – Testes de Hipóteses com o Excel

O Excel não oferece rotinas incorporadas para os testes de hipóteses apresentados neste capítulo. Para tratar dessas situações, apresentamos as planilhas do Excel que projetamos para testar hipóteses a respeito de uma média populacional e de uma proporção populacional. As planilhas são fáceis de usar e podem ser modificadas para manipular quaisquer dados amostrais. As planilhas estão disponíveis no seguinte endereço: http://thomsonlearning.com.br/estatapl.htm.

Média da População: σ Conhecido

Ilustramos nossa exposição usando o exemplo da distância percorrida pela bola de golfe MaxFlight apresentado na Seção 9.3. Os dados estão na coluna A de uma planilha do Excel. Consideramos que o desvio padrão populacional $\sigma = 12$ seja conhecido e que o nível de significância seja $\alpha = 0{,}05$. As etapas a seguir podem ser usadas para testar a hipótese H_0: $\mu = 295$ contra H_a: $\mu \neq 295$.

ARQUIVO DA INTERNET
Hyp Sigma Known

Consulte a Figura 9.11 à medida que descrevermos o procedimento. A planilha em segundo plano exibe as células com as fórmulas utilizadas para calcular os resultados apresentados na planilha em primeiro plano. Os dados são inseridos nas células A2:A51. As etapas a seguir são necessárias para se usar o modelo (*template*) para esse conjunto de dados.

Etapa 1. Digite o intervalo de dados A2:A51 na célula de fórmula =CONT.NÚM na célula D4.
Etapa 2. Digite o intervalo de dados A2:A51 na célula de fórmula =MÉDIA na célula D5.
Etapa 3. Digite o desvio padrão populacional $\sigma = 12$ na célula D6.
Etapa 4. Digite o valor hipotético 295 relativo à média populacional na célula D8.

As fórmulas de célula restantes fornecerão automaticamente o erro padrão, o valor z da estatística de teste e três valores p. Uma vez que a hipótese alternativa ($\mu_0 \neq 295$) indica um teste bicaudal, o valor p (Bicaudal) na célula D15 é usado para se tomar a decisão de rejeição. Com o valor $p = 0{,}1255 > \alpha = 0{,}05$, a hipótese nula não pode ser rejeitada. Os valores p nas células D13 ou D14 seriam usados se as hipóteses envolvessem um teste unicaudal.

Esse modelo pode ser usado para se fazer os cálculos de teste de hipóteses de outras aplicações. Por exemplo, para realizar um teste de hipótese para um novo conjunto de dados, insira os novos dados amostrais na coluna A da planilha. Modifique as fórmulas contidas nas células D4 e D5 para que correspondam ao novo intervalo de dados. Digite o desvio padrão da população na célula D6 e o valor hipotético para a média populacional na célula D8 para obter os resultados. Se os novos dados amostrais já tiverem sido sintetizados, eles não precisam ser inseridos na planilha. Nesse caso, digite o tamanho da amostra na célula D4, a média amostral na célula D5, o desvio padrão da população na célula D6 e o valor hipotético da média populacional na célula D8 para obter os resultados. A planilha da Figura 9.11 está disponível no arquivo Hyp Sigma Known na página do livro na internet.

Média da População: σ Desconhecido

Ilustramos nossa exposição usando o exemplo da avaliação do Aeroporto Heathrow apresentado na Seção 9.4. Os dados estão na coluna A de uma planilha do Excel. O desvio padrão σ da população é desconhe-

* O Minitab fornece simultaneamente os resultados do teste de hipóteses e os resultados de estimação por intervalo. O usuário pode selecionar qualquer nível de confiança para a estimação por intervalo da média populacional: aqui, sugerimos um intervalo de confiança de 95%.

cido e será estimado por meio do desvio padrão s da amostra. O nível de significância é $\alpha = 0{,}05$. Os passos a seguir podem ser usados para testar a hipótese H_0: $\mu \leq 7$ contra H_a: $\mu > 7$.

ARQUIVO DA INTERNET

Consulte a Figura 9.12 à medida que descrevermos o procedimento. A planilha em segundo plano exibe as fórmulas contidas em células que são usadas para calcular os resultados apresentados na versão da planilha em primeiro plano. Os dados são inseridos nas células A2:A61. As etapas a seguir são necessárias para se usar o modelo para esse conjunto de dados.

Figura 9.11 Planilha do Excel para testes de hipótese a respeito de uma média populacional para o caso em que σ é conhecido

Nota: As linhas 18 a 48 estão ocultas.

	A	B	C	D	E
1	Yards		Teste de Hipótese a Respeito de uma Média Populacional		
2	303		para o Caso em que σ é Conhecido		
3	282				
4	289		Tamanho da Amostra	=CONT.NÚM(A2:A51)	
5	298		Média Amostral	=MÉDIA(A2:A51)	
6	283		Desvio Padrão da Popul.	12	
7	317				
8	297		Valor Hipotético	295	
9	308				
10	317		Erro Padrão	=D6/RAIZ(D4)	
11	293		Estatística de Teste z	=(D5-D8)/D10	
12	284				
13	290		Valor p (Cauda Inferior)	=DIST.NORM(D11)	
14	304		Valor p (Cauda Superior)	=1-D13	
15	290		Valor p (Bicaudal)	=2*MÍNIMO(D13;D14)	
16	311				
17	305				
49	303				
50	301				
51	292				
52					

	A	B	C	D	E
1	Yards		Teste de Hipótese a Respeito de uma Média		
2	303		Populacional para o Caso em que σ é Conhecido		
3	282				
4	289		Tamanho da Amostra	50	
5	298		Média Amostral	297,6	
6	283		Desvio Padrão da Popul.	12	
7	317				
8	297		Valor Hipotético	295	
9	308				
10	317		Erro Padrão	1,70	
11	293		Estatística de Teste z	1,53	
12	284				
13	290		Valor p (Cauda Inferior)	0,9372	
14	304		Valor p (Cauda Superior)	0,0628	
15	290		Valor p (Bicaudal)	0,1255	
16	311				
17	305				
49	303				
50	301				
51	292				
52					

Etapa 1. Digite o intervalo de dados A2:A61 na célula de fórmula =CONT.NÚM na célula D4.
Etapa 2. Digite o intervalo de dados A2:A61 na célula de fórmula =MÉDIA na célula D5.
Etapa 3. Digite o intervalo de dados A2:A61 na célula de fórmula =DESVPAD na célula D6.
Etapa 4. Digite o valor hipotético 7 relativo à média populacional na célula D8

As fórmulas de célula restantes fornecerão automaticamente o erro padrão, o valor t da estatística de teste, o número de graus de liberdade e três valores p. Uma vez que a hipótese alternativa ($\mu > 7$) indica um teste da cauda superior, o valor p (Cauda Superior) na célula D15 é usado para se tomar a decisão.

Figura 9.12 Planilha do Excel para testes de hipóteses a respeito de uma média populacional para o caso em que σ é desconhecido

	A	B	C	D	E
1	Rating		Teste de Hipóteses a Respeito de uma Média Populacional		
2	5		para o Caso em que σ é Desconhecido		
3	7				
4	8		Tamanho da Amostra	=CONT.NÚM(A2:A61)	
5	7		Média Amostral	=MÉDIA(A2:A61)	
6	8		Desvio Padrão da Popul.	=DESVPAD(A2:A61)	
7	8				
8	8		Valor Hipotético	7	
9	7				
10	8		Erro Padrão	=D6/RAIZ(D4)	
11	10		Estatística de Teste t	=(D5-D8)/D10	
12	6		Graus de Liberdade	=D4-1	
13	7				
14	8		Valor p (Cauda Inferior)	=SE(D11,0,DISTT(-D11,D12,1),1-DISTT(D11,D12,1)	
15	8		Valor p (Cauda Superior)	=1-D14	
16	9		Valor p (Bicaudal)	=2*MÍNIMO(D14,D15)	
17	7				
59	7				
60	7				
61	8				
62					

	A	B	C	D	E
1	Rating		Teste de Hipóteses a Respeito de uma Média		
2	5		Populacional o Caso em que σ é Desconhecido		
3	7				
4	8		Tamanho da Amostra	60	
5	7		Média Amostral	7,25	
6	8		Desvio Padrão da Popul.	1,05	
7	8				
8	8		Valor Hipotético	7	
9	7				
10	8		Erro Padrão	0,136	
11	10		Estatística de Teste t	1,841	
12	6		Graus de Liberdade	59	
13	7				
14	8		Valor p (Cauda Inferior)	0,9647	
15	8		Valor p (Cauda Superior)	0,0353	
16	9		Valor p (Bicaudal)	0,0706	
17	7				
59	7				
60	7				
61	8				
62					

Nota: As linhas 18 a 58 estão ocultas.

Com o valor $p = 0{,}353 < \alpha = 0{,}05$, a hipótese nula é rejeitada. Os valores p nas células D14 ou D16 seriam usados se as hipóteses envolvessem um teste da cauda inferior ou um teste bicaudal.

Esse modelo pode ser usado para se fazer os cálculos de teste de hipóteses de outras aplicações. Por exemplo, para realizar um teste de hipóteses para um novo conjunto de dados, insira os novos dados amostrais na coluna A da planilha e modifique as fórmulas contidas nas células D4 e D5 e D6 para que correspondam ao novo intervalo de dados. Digite o valor hipotético da média populacional na célula D8 para obter os resultados. Se os novos dados amostrais já tiverem sido sintetizados, eles não precisam ser inseridos na planilha.

Nesse caso, digite o tamanho da amostra na célula D4, a média amostral na célula D5, o desvio padrão da amostra na célula D6 e o valor hipotético da média populacional na célula D8 para obter os resultados. A planilha da Figura 9.12 está disponível no arquivo Hyp Sigma Unknown na internet.

Proporção da População

ARQUIVO DA INTERNET

Hypothesis p

Ilustramos nossa exposição usando os dados da pesquisa do curso de golfe Pine Creek apresentados na Seção 9.5. Os dados dos golfistas Homem ou Mulher estão na coluna A de uma planilha do Excel. Consulte a Figura 9.13 à medida que descrevermos o procedimento. A planilha em segundo plano exibe as células com as fórmulas utilizadas para calcular os resultados apresentados na planilha em primeiro plano.

Figura 9.13 Planilha do Excel para testes de hipóteses a respeito de uma proporção populacional

	A	B	C	D	E
1	Golfista		Estimação por Intervalo de uma Proporção Populacional		
2	Mulher				
3	Homem		Tamanho da Amostra	=CONT.VALORES(A2:A401)	
4	Mulher		Resposta de Interesse	Mulher	
5	Homem		Contagem da Resposta	=CONT.SE(A2;A401;D4)	
6	Homem		Proporção da Amostra	=D5/D3	
7	Mulher				
8	Homem		Valor Hipotético	0,20	
9	Homem				
10	Mulher		Erro Padrão	=RAIZ(D8*(1-D8)/D3)	
11	Homem		Estatística de Teste *z*	=(D6-D8)/D10	
12	Homem				
13	Homem		Valor *p* (Cauda Inferior)	=DIST.NORM(D11)	
14	Homem		Valor *p* (Cauda Superior)	=1-D13	
15	Homem		Valor *p* (Bicaudal)	=2*MÍNIMO(D13;D14)	
16	Mulher				
400	Homem				
401	Homem				
402					

	A	B	C	D	E
1	Golfista		Estimação por Intervalo de uma Proporção Populacional		
2	Mulher				
3	Homem		Tamanho da Amostra	400	
4	Mulher		Resposta de Interesse	Mulher	
5	Homem		Contagem da Resposta	100	
6	Homem		Proporção da Amostra	0,2500	
7	Mulher				
8	Homem		Valor Hipotético	0,20	
9	Homem				
10	Mulher		Erro Padrão	0,0200	
11	Homem		Estatística de Teste *z*	2,50	
12	Homem				
13	Homem		Valor *p* (Cauda Inferior)	0,9938	
14	Homem		Valor *p* (Cauda Superior)	0,0062	
15	Homem		Valor *p* (Bicaudal)	0,0124	
16	Mulher				
400	Homem				
401	Homem				
402					

Nota: As linhas 17 a 399 estão ocultas.

Os dados estão inseridos nas células A2:A401. As etapas a seguir podem ser usadas para testar a hipótese H_0: $p \leq 0{,}20$ contra H_a: $p > 0{,}20$.

Etapa 1. Digite o intervalo de dados A2:A401 na célula de fórmula =CONT.VALORES na célula D3.
Etapa 2. Digite Mulher como a resposta de interesse na célula D4.
Etapa 3. Digite o intervalo de dados A2:A401 na célula de fórmula =CONT.SE na célula D5.
Etapa 4. Digite o valor hipotético 0,20 relativo à proporção populacional na célula D8.

As fórmulas de célula restantes fornecerão automaticamente o erro padrão, o valor z da estatística de teste e três valores p. Uma vez que a hipótese alternativa ($p_0 > 0{,}20$) indica um teste da cauda superior, o valor p (Cauda Superior) na célula D14 é usado para se tomar a decisão. Com o valor $p = 0{,}0062 < \alpha = 0{,}05$, a hipótese nula é rejeitada. Os valores p nas células D13 ou D15 seriam usados se a hipótese envolvesse um teste da cauda inferior ou um teste bicaudal.

Esse modelo pode ser usado para se fazer os cálculos de teste de hipóteses de outras aplicações. Por exemplo, para realizar um teste de hipóteses para um novo conjunto de dados, insira os novos dados amostrais na coluna A da planilha. Modifique as fórmulas contidas nas células D3 e D5 para que correspondam ao novo intervalo de dados. Digite a resposta de interesse na célula D4 e o valor hipotético da proporção populacional na célula D8 para obter os resultados. Se os novos dados amostrais já tiverem sido sintetizados, eles não precisam ser inseridos na planilha. Nesse caso, digite o tamanho da amostra na célula D3, a proporção amostral na célula D6 e o valor hipotético da proporção populacional na célula D8 para obter os resultados. A planilha da Figura 9.13 está disponível no arquivo Hypothesis p na internet.

Comparações Envolvendo Médias

A ESTATÍSTICA NA PRÁTICA

FISONS CORPORATION
Rochester, NY

A Fisons Corporation de Rochester, no estado de Nova York, é uma unidade da Fisons Plc., do Reino Unido. A empresa iniciou suas operações nos Estados Unidos em 1966.

A Fisons Pharmaceutical Division utiliza amplos procedimentos estatísticos para testar e desenvolver novos medicamentos. O processo de testes na indústria farmacêutica geralmente é composto por três etapas: (1) testes pré-clínicos, (2) testes de uso e de segurança em longo prazo e (3) testes da eficácia clínica. Em cada etapa sucessiva, decresce a chance de um medicamento ser aprovado nos rigorosos testes; entretanto, o custo da realização de testes adicionais se eleva drasticamente. Levantamentos realizados pela indústria indicam que o processo de pesquisa e desenvolvimento de um novo medicamento custa, em média, US$ 250 milhões e demanda 12 anos. Portanto, é importante eliminar novos medicamentos malsucedidos nas fases iniciais do processo de testes, bem como identificar aqueles que são promissores para serem submetidos a testes adicionais.

A estatística desempenha papel importante nas pesquisas farmacêuticas, sendo uma área em que o controle governamental é severo e aplicado com rigor. Nos testes pré-clínicos, estudos estatísticos de duas ou três populações normalmente são utilizados para determinar se o novo medicamento deve continuar a ser estudado no programa de uso e de segurança em longo prazo. As populações podem consistir no novo medicamento, no controle e no medicamento padrão. O processo de testes pré-clínicos inicia-se quando um novo medicamento é enviado à equipe de Farmacologia para avaliação de sua eficácia – a capacidade de o medicamento produzir os efeitos desejados. Como parte do processo, um estatístico é solicitado a projetar um expe-

rimento que possa ser usado para testar a nova droga. O projeto deve especificar o tamanho da amostra e os métodos estatísticos de análise. Em um estudo de duas populações, uma amostra é usada para se obterem dados sobre a eficácia do novo medicamento (população 1) e uma segunda amostra é utilizada para se obterem dados sobre a eficácia de um medicamento padrão (população 2). Dependendo da utilização pretendida, o novo medicamento e o medicamento padrão são testados em diversas áreas, como neurologia, cardiologia e imunologia. Na maioria dos estudos, o método estatístico envolve o teste de hipóteses quanto à diferença entre as médias da população do novo medicamento e a população do medicamento padrão. Se faltar eficácia a um novo medicamento ou se ele produzir efeitos indesejáveis em comparação com o medicamento padrão, esse novo medicamento será rejeitado e eliminado dos testes adicionais. Somente os novos medicamentos que apresentam comparações promissoras em relação aos medicamentos padrão são encaminhados ao programa de testes de uso e de segurança em longo prazo.

A coleta de dados adicionais e estudos de populações múltiplas são realizados no programa de testes de uso e de segurança em longo prazo e nos programas de testes clínicos. A Food and Drug Administration (FDA) exige que os métodos estatísticos sejam definidos antes da realização desses testes para evitar distorções relacionadas aos dados. Além disso, para evitar vieses humanos, alguns dos ensaios clínicos são realizados com o método de duplo ou triplo-cego. Ou seja, nem o sujeito nem o investigador sabem qual medicamento é administrado a quem. Se o novo medicamento cumprir todas as exigências em relação ao medicamento padrão, o pedido de registro de uma nova droga é feito na FDA. O pedido de registro é examinado rigorosamente por estatísticos e cientistas do departamento.

Neste capítulo, você aprenderá a construir estimações por intervalo e a fazer testes de hipótese a respeito de médias e proporções com duas populações. Serão apresentadas técnicas para analisar amostras aleatórias independentes, bem como amostras relacionadas.

Nos Capítulos 8 e 9, mostramos como desenvolver estimações por intervalo e como realizar testes de hipótese para situações que envolvem uma média populacional. Neste capítulo, estendemos nossa discussão da inferência estatística a aplicações que comparam as médias de duas ou mais populações. Por exemplo, talvez queiramos desenvolver uma estimação por intervalo da diferença entre a média dos salários iniciais de uma população de homens e a média dos salários iniciais de uma população de mulheres, ou testar a hipótese de que o número médio de horas entre a ocorrência de panes é o mesmo para quatro diferentes máquinas. Iniciamos mostrando como desenvolver estimações por intervalo e realizar testes de hipóteses a respeito da diferença entre duas médias populacionais, quando se presume que dois desvios padrão populacionais sejam conhecidos.

10.1 INFERÊNCIAS SOBRE A DIFERENÇA ENTRE AS MÉDIAS DE DUAS POPULAÇÕES: σ_1 E σ_2 CONHECIDOS

Admitindo que μ_1 denota a média da população 1 e μ_2, a média da população 2, vamos nos concentrar nas inferências sobre a diferença entre as médias: $\mu_1 - \mu_2$. Para fazermos uma inferência sobre essa diferença, selecionamos uma amostra aleatória simples de n_1 unidades da população 1 e uma amostra aleatória simples de n_2 unidades da população 2. As duas amostras, tomadas separada e independentemente, são chamadas **amostras aleatórias simples independentes**. Nesta seção, presumiremos que existam informações disponíveis, de tal forma que é possível supor que os desvios padrão, σ_1 e σ_2, das duas populações sejam conhecidos antes de se coletarem as amostras. Referimo-nos a essa situação como o caso em que σ_1 e σ_2 são conhecidos. No exemplo a seguir, vamos mostrar como calcular a margem de erro e desenvolver uma estimação por intervalo da diferença entre as duas médias populacionais.

Estimação por Intervalo de $\mu_1 - \mu_2$

A Greystone Department Stores, Inc. opera duas lojas em Buffalo, no estado de Nova York: uma no centro da cidade e a outra em um *shopping center* da periferia. O gerente regional notou que os produtos que têm boa vendagem em uma loja nem sempre vendem bem na outra. O gerente acredita que essa situação talvez se deva a diferenças nos aspectos demográficos entre os clientes das duas localidades. Os clientes talvez difiram em termos de idade, nível educacional, renda etc. Suponha que o gerente nos peça para investigar a diferença entre as médias de idade dos clientes que compram nas duas lojas.

Vamos definir a população 1 como todos os clientes que compram na loja do centro da cidade e a população 2 como todos os clientes que compram na loja da periferia.

μ_1 = média da população 1 (ou seja, a média de idade de todos os clientes que compram na loja do centro da cidade).

μ_2 = média da população 2 (ou seja, a média de idade de todos os clientes que compram na loja da periferia).

A diferença entre as médias das duas populações é $\mu_1 - \mu_2$.

Para estimar $\mu_1 - \mu_2$, selecionamos uma amostra aleatória simples de n_1 clientes da população 1 e uma amostra aleatória simples de n_2 clientes da população 2. Então, calculamos as duas médias amostrais:

$\bar{x}_1$ = média amostral da idade de uma amostra aleatória simples de n_1 clientes do centro da cidade.

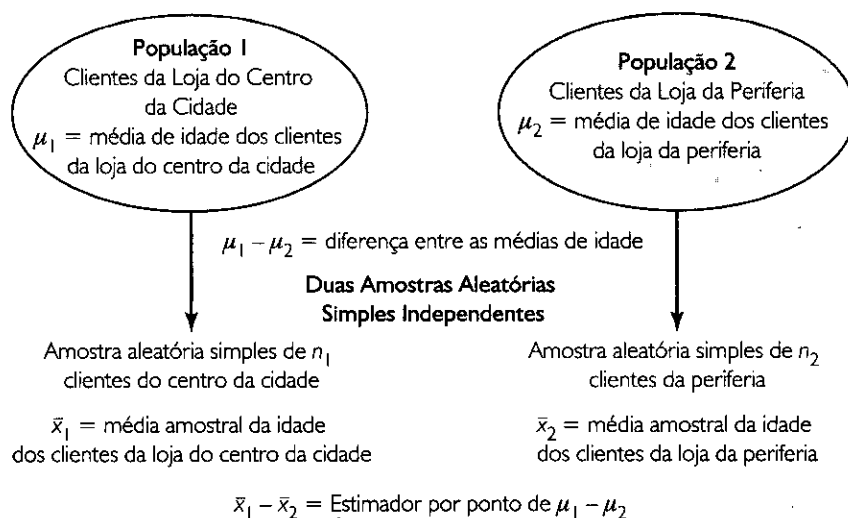
$\bar{x}_2$ = média amostral da idade de uma amostra aleatória simples de n_2 clientes da periferia.

O estimador por ponto da diferença entre as médias das duas populações é a diferença entre as duas médias amostrais.

ESTIMADOR POR PONTO DA DIFERENÇA ENTRE AS MÉDIAS DE DUAS POPULAÇÕES

$$\bar{x}_1 - \bar{x}_2 \qquad \textbf{(10.1)}$$

A Figura 10.1 apresenta uma visão geral do processo utilizado para estimar a diferença entre as duas médias populacionais baseadas em duas amostras aleatórias simples independentes.

Figura 10.1 Estimando a diferença entre as médias de duas populações

À semelhança do que ocorre com outros estimadores por ponto, o estimador por ponto de $\bar{x}_1 - \bar{x}_2$ tem um erro padrão que descreve a variação da distribuição amostral do estimador. Com duas amostras aleatórias independentes, o erro padrão de $\bar{x}_1 - \bar{x}_2$ é o seguinte:

O erro padrão de $\bar{x}_1 - \bar{x}_2$ é o desvio padrão da distribuição amostral de $\bar{x}_1 - \bar{x}_2$.

$$\sigma_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \qquad \textbf{(10.2)}$$

Se ambas as populações tiverem uma distribuição normal, ou se os tamanhos de amostra forem suficientemente grandes a ponto de o teorema do limite central nos permitir concluir que as distribuições amostrais de $\bar{x}_1$ e de $\bar{x}_2$ possam ser aproximadas a uma distribuição normal, a distribuição amostral de $\bar{x}_1 - \bar{x}_2$ terá uma distribuição normal com uma média dada por $\mu_1 - \mu_2$.

Conforme mostramos no Capítulo 8, uma estimação por intervalo é obtida por uma estimação por ponto ± uma margem de erro. No caso da estimação da diferença entre duas médias populacionais, uma estimação por intervalo assumirá a seguinte forma:

$$\bar{x}_1 - \bar{x}_2 \pm \text{Margem de erro}$$

Com a distribuição amostral de $\bar{x}_1 - \bar{x}_2$ tendo uma distribuição normal, podemos escrever a margem de erro da seguinte maneira:

A margem de erro é obtida multiplicando-se o erro padrão por $z_{\alpha/2}$.

$$\text{Margem de erro} = z_{\alpha/2}\sigma_{\bar{x}_1-\bar{x}_2} = z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \qquad \textbf{(10.3)}$$

Desse modo, a estimação por intervalo da diferença entre as duas médias populacionais é a seguinte:

ESTIMAÇÃO POR INTERVALO DA DIFERENÇA ENTRE AS MÉDIAS DE DUAS POPULAÇÕES: σ_1 E σ_1 CONHECIDOS

$$\bar{x}_1 - \bar{x}_2 \pm z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \qquad \textbf{(10.4)}$$

em que $1 - \alpha$ é o coeficiente de confiança.

Retornemos ao exemplo da Greystone. Com base em estudos anteriores sobre os aspectos demográficos dos clientes, sabe-se que os desvios padrão das duas populações são $\sigma_1 = 9$ anos e $\sigma_2 = 10$ anos, respectivamente. Os dados coletados de duas amostras aleatórias simples independentes de clientes da Greystone forneceram os seguintes resultados.

	Loja do Centro da Cidade	Loja da Periferia
Tamanho da Amostra	$\bar{n}_1 = 36$	$\bar{n}_2 = 49$
Média Amostral	$x_1 = 40$ anos	$x_2 = 35$ anos

Usando a Equação 10.1, descobrimos que a estimação por ponto da diferença entre a média de idade das duas populações é $\bar{x}_1 - \bar{x}_2 = 40 - 35 = 5$ anos. Assim, calculamos que os clientes da loja do centro da cidade têm uma média de idade 5 anos maior que a média de idade dos clientes da loja da periferia. Agora, podemos usar a Equação 10.4 para calcular a margem de erro e produzir a estimação por intervalo de $\mu_1 - \mu_2$. Usando 95% de confiança e $z_{\alpha/2} = z_{0,025} = 1,96$, obtemos:

$$\bar{x}_1 - \bar{x}_2 \pm z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

$$40 - 35 \pm 1,96\sqrt{\frac{9^2}{36} + \frac{10^2}{49}}$$

$$5 \pm 4,06$$

Desse modo, a margem de erro é 4,06 anos e a estimação por intervalo de confiança de 95% da diferença entre as duas médias populacionais é $5 - 4,06 = 0,94$ anos a $5 + 4,06 = 9,06$ anos.

Testes de Hipóteses sobre $\mu_1 - \mu_2$

Consideremos os testes de hipóteses sobre a diferença entre as médias de duas populações. Usando D_0 para denotar as diferenças hipotéticas entre μ_1 e μ_2, as três formas de um teste de hipóteses são as seguintes:

$$H_0: \mu_1 - \mu_2 \geq D_0 \qquad H_0: \mu_1 - \mu_2 \leq D_0 \qquad H_0: \mu_1 - \mu_2 = D_0$$
$$H_\alpha: \mu_1 - \mu_2 < D_0 \qquad H_\alpha: \mu_1 - \mu_2 > D_0 \qquad H_\alpha: \mu_1 - \mu - \neq D_0$$

Em muitas aplicações, $D_0 = 0$. Usando o teste bicaudal como exemplo, quando $D_0 = 0$, a hipótese nula é H_0: $\mu_1 - \mu_2 = 0$. Nesse caso, a hipótese nula é que μ_1 e μ_2 são iguais. A rejeição de H_0 leva à conclusão de que H_α: $\mu_1 - \mu_2 \neq 0$ é verdadeira; ou seja, μ_1 e μ_2 não são iguais.

As etapas para realizar os testes de hipótese apresentados no Capítulo 9 são aplicáveis aqui. Precisamos escolher um nível de significância, calcular o valor da estatística de teste e encontrar o valor p para determinar se a hipótese nula deve ser rejeitada. Com duas amostras aleatórias simples independentes, mostramos que o estimador por ponto $\bar{x}_1 - \bar{x}_2$ tem o erro padrão $\sigma_{\bar{x}_1-\bar{x}_2}$, dado pela Equação 10.2, e que a distribuição de $\bar{x}_1 - \bar{x}_2$ pode ser descrita por uma distribuição normal. Nesse caso, a estatística de teste da diferença entre as duas médias populacionais quando σ_1 e σ_2 são conhecidos é a seguinte:

ESTATÍSTICA DE TESTE PARA TESTES DE HIPÓTESES SOBRE $\mu_1 - \mu_2$ QUANDO σ_1 E σ_2 SÃO CONHECIDOS

$$z = \frac{(\bar{x}_1 - \bar{x}_2) - D_0}{\sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}} \qquad \textbf{(10.5)}$$

Vamos demonstrar o uso dessa estatística de teste no seguinte exemplo de teste de hipóteses.

Como parte de um estudo para avaliar as diferenças na qualidade educacional entre dois centros de ensino, um exame padronizado é aplicado a pessoas que estudam nesses centros. A diferença entre a média das notas obtidas no exame é usada para avaliar as diferenças de qualidade entre os centros. As médias populacionais correspondentes aos dois centros são as seguintes:

μ_1 = a média das notas de exame da população de pessoas que estudam no centro A.

μ_2 = a média das notas de exame da população de pessoas que estudam no centro B.

Iniciamos com a hipótese experimental de que não existe diferença entre a qualidade de ensino ministrado nos dois centros. Portanto, em termos da média das notas de exame, a hipótese nula é que $\mu_1 - \mu_2 = 0$. Se as evidências amostrais levarem à rejeição dessa hipótese, concluiremos que a média das notas de exame diferem com respeito às duas populações. Essa conclusão indica um diferencial de qualidade entre os dois centros e sugere que talvez seja necessário um estudo de acompanhamento para investigar a razão desse diferencial. As hipóteses nula e alternativa desse teste bicaudal são as seguintes:

ARQUIVO DA INTERNET
ExamScores

$$H_0: \mu_1 - \mu_2 = 0$$
$$H_\alpha: \mu_1 - \mu_2 \neq 0$$

O exame padronizado, aplicado anteriormente em uma série de ambientes educacionais, sempre resultou em um desvio padrão de notas de exame próximo a 10 pontos. Desse modo, usaremos essa informação para supor que os desvios padrão populacionais sejam conhecidos, sendo $\sigma_1 = 10$ e $\sigma_2 = 10$. Um nível de significância $\alpha = 0{,}05$ é especificado para o estudo.

São tomadas amostras aleatórias simples independentes de $n_1 = 30$ indivíduos do centro de ensino A e $n_2 = 40$ indivíduos do centro de ensino B. As respectivas médias amostrais são $\bar{x}_1 = 82$ e $\bar{x}_2 = 78$. Esses dados sugerem uma diferença significativa entre as médias populacionais dos dois centros de ensino? Para ajudar a responder a essa pergunta, calculamos a estatística de teste usando a Equação 10.5:

$$z = \frac{(\bar{x}_1 - \bar{x}_2) - D_0}{\sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}} = \frac{(82 - 78) - 0}{\sqrt{\dfrac{10^2}{30} + \dfrac{10^2}{40}}} = 1{,}66$$

Calculemos, agora, o valor p para esse teste bicaudal. Uma vez que a estatística de teste z está na cauda superior, calculamos primeiramente a área sob a curva à direita de $z = 1{,}66$. Usando a tabela de distribuição normal padrão, verificamos que a área entre a média e $z = 1{,}66$ é 0,4515. Desse modo, a área da cauda superior da distribuição é $0{,}5000 - 0{,}4515 = 0{,}0485$. Desde que este seja um teste bicaudal, devemos duplicar a área da cauda: valor $p = 2(0{,}485) = 0{,}0970$. Ao seguir a regra habitual de rejeitar H_0 se o valor $p \leq \alpha$, vemos que o valor p igual a 0,970 não nos permite rejeitar H_0 ao nível de significância 0,05. Os resultados amostrais não fornecem evidências suficientes para concluirmos que os centros de ensino diferem em termos de qualidade.

Neste capítulo, usaremos o critério do valor p para o teste de hipóteses, conforme descrevemos no Capítulo 9. Entretanto, se você preferir, a estatística de teste e a regra de rejeição pelo valor crítico podem ser usadas. Com $\alpha = 0{,}05$ e $z_{\alpha/2} = z_{0{,}025} = 1{,}96$, a regra de rejeição empregando-se o critério do valor crítico seria rejeitar H_0 se $z \leq -1{,}96$ ou se $z \geq 1{,}96$. Com $z = 1{,}66$, chegamos à mesma conclusão de não rejeitar H_0.

No exemplo anterior, demonstramos um teste de hipóteses bicaudal a respeito da diferença entre duas médias populacionais. Testes da cauda inferior e da cauda superior também podem ser considerados. Esses testes usam a mesma estatística de teste apresentada na Equação 10.5. O procedimento para calcular o valor p e a regra de rejeição para esses testes bicaudais são idênticos aos apresentados no Capítulo 9.

Conselho Prático

AUTOTESTE

Na maioria das aplicações dos procedimentos de estimação por intervalo e de teste de hipóteses apresentados nesta seção, variáveis aleatórias com $n_1 \geq 30$ e $n_2 \geq 30$ são adequadas. Nos casos em que um ou outro tamanho de amostra, ou ambos, forem menores que 30, as distribuições das populações tornam-se considerações importantes. Em geral, com tamanhos de amostra menores, é mais importante que o analista se convença de que é razoável presumir que as distribuições das duas populações sejam, no mínimo, aproximadamente normais.

Exercícios

Métodos

AUTOTESTE

1. Considere os seguintes resultados, referentes a duas amostras aleatórias independentes tomadas de duas populações:

Amostra 1	**Amostra 2**
$n_1 = 50$	$n_2 = 35$
$\bar{x}_1 = 13{,}6$	$\bar{x}_2 = 11{,}6$
$\sigma_1 = 2{,}2$	$\sigma_2 = 3{,}0$

a. Qual é a estimação por ponto da diferença entre as duas médias populacionais?
b. Apresente um intervalo de confiança de 90% relativo à diferença entre as duas médias populacionais.
c. Apresente um intervalo de confiança de 95% relativo à diferença entre as duas médias populacionais.

2. Considere o seguinte teste de hipóteses:

$$H_0: \mu_1 - \mu_2 \leq 0$$
$$H_\alpha: \mu_1 - \mu_2 > 0$$

Os resultados a seguir referem-se a duas amostras independentes tomadas de duas populações.

Amostra 1	**Amostra 2**
$n_1 = 40$	$n_2 = 50$
$\bar{x}_1 = 25{,}2$	$\bar{x}_2 = 22{,}8$
$\sigma_1 = 5{,}2$	$\sigma_2 = 6{,}0$

a. Qual é o valor da estatística de teste?
b. Qual é o valor p?
c. Com $\alpha = 0{,}05$, qual é a conclusão do seu teste de hipóteses?

3. Considere o seguinte teste de hipóteses:

$$H_0: \mu_1 - \mu_2 = 0$$
$$H_\alpha: \mu_1 - \mu_2 \neq 0$$

Os resultados a seguir referem-se a duas amostras independentes tomadas de duas populações:

Amostra 1	**Amostra 2**
$n_1 = 80$	$n_2 = 70$
$\bar{x}_1 = 104$	$\bar{x}_2 = 106$
$\sigma_1 = 8{,}4$	$\sigma_2 = 7{,}6$

a. Qual é o valor da estatística de teste?
b. Qual é o valor p?
c. Com $\alpha = 0{,}05$, qual é a conclusão do seu teste de hipóteses?

Aplicações

4. A alta dos preços da gasolina atingiu níveis recordes em 16 estados durante 2003 (*The Wall Street Journal*, 7 de março de 2003). Dois dos estados afetados foram a Califórnia e a Flórida. A American Automobile Association relatou um preço médio amostral de US$ 2,04 por galão (US$ 0,54 por litro) na Califórnia e um preço médio amostral de US$ 1,72 por galão (US$ 0,46 por litro) na Flórida. Use um tamanho de amostra 40 para os dados da Califórnia e um tamanho de amostra 35 para os dados da Flórida. Suponha que estudos anteriores indicando os desvios padrão populacionais de 0,10 para a Califórnia e de 0,08 para a Flórida sejam razoáveis.
   a. Qual é a estimação por ponto da diferença entre os preços médios populacionais por galão na Califórnia e na Flórida?
   b. Com 95% de confiança, qual é a margem de erro?
   c. Qual é a estimação por intervalo de confiança de 95% da diferença entre os preços médios populacionais por galão nos dois estados?

5. Um estudo realizado pela Cornell University dos diferenciais de salário entre homens e mulheres relatou que uma das razões pelas quais os salários dos homens são mais altos que os das mulheres é o fato de os homens tenderem a ter mais anos de experiência no trabalho que as mulheres (*Business Week*, 28 de agosto de 2000). Suponha que os seguintes resumos amostrais apresentem os anos de experiência correspondentes a cada grupo:

Homens	**Mulheres**
$n_1 = 100$	$n_2 = 85$
$\bar{x}_1 = 14{,}9$ anos	$\bar{x}_2 = 10{,}3$ anos
$\sigma_1 = 5{,}2$ anos	$\sigma_2 = 3{,}8$ anos

   a. Qual é a estimação por ponto da diferença entre as duas médias populacionais?
   b. Com 95% de confiança, qual é a margem de erro?
   c. Qual é a estimação por intervalo de confiança da diferença entre as duas médias populacionais?

6. As 40 mil corretoras imobiliárias do país estão entre os pequenos negócios mais lucrativos nos Estados Unidos. Essas empresas de baixo-perfil encontram empréstimos para os clientes em troca de comissões. A Mortgage Bankers Association of America divulga dados sobre o tamanho médio dos empréstimos manuseados pelas corretoras imobiliárias (*The Wall Street Journal*, 24 de fevereiro de 2003). Usando dados amostrais coerentes com os dados da Mortgage Bankers Association, uma amostra de 270 empréstimos realizados em 2002 forneceu um valor médio de empréstimos de US$ 175 mil. Dados de 2001 apresentaram uma amostra de 250 empréstimos realizados, com um valor médio de empréstimos de US$ 165 mil. Com base nos dados históricos dos empréstimos, pode-se presumir que os desvios padrão populacionais sejam conhecidos, sendo US$ 55 mil em 2002 e US$ 50 mil em 2001. Os dados amostrais indicam um aumento do valor médio de empréstimo entre 2001 e 2002? Use $\alpha = 0{,}05$.

7. Durante a temporada de 2003, a Major League Baseball tomou medidas para aumentar a velocidade de jogo nos jogos de beisebol a fim de manter o interesse da torcida (*CNN Headline News*, 30 de setembro de 2003). Os resultados apresentados a seguir são de uma amostra de 60 jogos disputados durante o verão de 2002 e de uma amostra de 50 jogos disputados durante o verão de 2003. A média amostral exibe a duração média dos jogos incluídos em cada amostra.

Temporada de 2002	**Temporada de 2003**
$n_1 = 60$	$n_2 = 50$
$\bar{x}_1 = 2$ horas e 52 minutos	$\bar{x}_2 = 2$ horas e 46 minutos

   a. Uma hipótese de pesquisa era que as medidas tomadas durante a temporada de 2003 reduziriam a duração média da população de jogos de beisebol. Formule as hipóteses nula e alternativa.
   b. Qual é a estimação por ponto da redução da duração média dos jogos na temporada de 2003?
   c. Dados históricos indicam que um desvio padrão populacional de 12 minutos é uma suposição razoável para ambos os anos. Realize um teste de hipóteses e relate qual é o valor p. Com o nível de significância 0,05, qual é a sua conclusão?
   d. Forneça uma estimação por intervalo de confiança de 95% sobre a redução da duração média dos jogos na temporada de 2003.
   e. Qual foi a redução percentual da média de tempo dos jogos de beisebol durante a temporada de 2003? A administração deve estar satisfeita com os resultados da análise estatística? Discuta o assunto. A duração dos jogos de beisebol deve continuar a ser uma preocupação no futuro? Explique.

8. Arnold Palmer e Tiger Woods são dois dos melhores golfistas da história desse esporte. Para mostrar como esses dois golfistas se comparariam se ambos estivessem jogando em sua melhor forma, os seguintes dados amostrais apresentam os resultados de suas pontuações em 18 buracos durante um campeonato promovido pela PGA. As pontuações de Palmer referem-se ao que ele obteve em sua temporada de 1960, enquanto as pontuações de Woods são de sua temporada de 1999 (*Golf Magazine*, fevereiro de 2000).

Arnold Palmer	Tiger Woods
$n_1 = 112$	$n_2 = 84$
$\bar{x}_1 = 69{,}95$	$\bar{x}_2 = 69{,}56$

Use os dados amostrais para testar a hipótese de que não existe diferença entre a média populacional de pontuações nos 18 buracos para os dois golfistas.

a. Suponha um desvio padrão populacional igual a 2,5 para ambos os golfistas. Qual é o valor da estatística de teste?
b. Qual é o valor p?
c. Com $\alpha = 0{,}01$, qual é a sua conclusão?

10.2 INFERÊNCIAS SOBRE A DIFERENÇA ENTRE AS MÉDIAS DE DUAS POPULAÇÕES: σ_1 E σ_2 DESCONHECIDOS

Nesta seção, estendemos a discussão das inferências sobre a diferença entre duas médias populacionais para o caso em que os dois desvios padrão, σ_1 e σ_2, são desconhecidos. Nesse caso, usaremos os desvios padrão amostrais, s_1 e s_2, para estimar os desvios padrão populacionais desconhecidos. Quando utilizarmos os desvios padrão amostrais, os procedimentos de estimação por intervalo e de teste de hipóteses vão se basear na distribuição t em vez da distribuição normal padrão.

Estimação por Intervalo de $\mu_1 - \mu_2$

ARQUIVO DA INTERNET
CheckAcct

No exemplo que apresentamos a seguir, mostramos como calcular a margem de erro e como desenvolver uma estimação por intervalo da diferença entre duas médias populacionais, quando σ_1 e σ_2 são desconhecidos. O Clearwater National Bank realiza um estudo idealizado para identificar as diferenças na utilização das contas correntes pelos clientes em dois de seus bancos filiais.

Uma amostra aleatória simples de 28 contas correntes é selecionada da filial situada em Cherry Grove e uma amostra aleatória simples independente é selecionada de sua filial em Beechmont. O saldo atual da conta corrente é registrado para cada uma das contas. Apresentamos a seguir um resumo dos saldos bancários:

	Cherry Grove	Beechmont
Tamanho da amostra	$n_1 = 28$	$n_2 = 22$
Média amostral	x_1 = US\$ 1.025	x_2 = US\$ 910
Desvio Padrão da Amostra	s_1 = US\$ 150	s_2 = US\$ 125

O Clearwater National Bank quer estimar a diferença entre o saldo médio das contas correntes mantidas pela população de clientes de Cherry Grove e da população de clientes de Beechmont. Vamos desenvolver a margem de erro e uma estimação por intervalo da diferença entre essas duas médias populacionais.

Quando σ_1 e σ_2 são estimados por meio de s_1 e s_2, a distribuição t é usada para se fazer inferências sobre a diferença entre duas médias populacionais.

Na Seção 10.1, apresentamos a seguinte estimação por intervalo para o caso em que os desvios padrão populacionais, σ_1 e σ_2, são conhecidos.

$$\bar{x}_1 - \bar{x}_2 \pm z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

Sendo σ_1 e σ_2 desconhecidos, usaremos os desvios padrão amostrais s_1 e s_2 para estimar σ_1 e σ_2 e substituiremos $z_{\alpha/2}$ por $t_{\alpha/2}$. Conseqüentemente, a estimação por intervalo da diferença entre duas médias populacionais é dada pela seguinte expressão:

ESTIMAÇÃO POR INTERVALO DA DIFERENÇA ENTRE DUAS MÉDIAS POPULACIONAIS QUANDO σ_1 E σ_2 SÃO DESCONHECIDOS

$$\bar{x}_1 - \bar{x}_2 \pm t_{\alpha/2}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \qquad \textbf{(10.6)}$$

em que $1 - \alpha$ é o coeficiente de confiança.

Nessa expressão, o uso da distribuição t é uma aproximação, mas ela produz excelentes resultados e é relativamente fácil de usar. A única dificuldade que encontramos ao usar a Equação 10.6 é determinar os graus de liberdade apropriados para $t_{\alpha/2}$. Softwares estatísticos calculam automaticamente os graus de liberdade apropriados. A fórmula usada é a seguinte:

$$gl = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{1}{n_1 - 1}\left(\frac{s_1^2}{n_1}\right)^2 + \frac{1}{n_2 - 1}\left(\frac{s_2^2}{n_2}\right)^2} \qquad \textbf{(10.7)}$$

Vamos retornar ao exemplo do Clearwater National Bank e mostrar como usar a Equação 10.6 para produzir uma estimação por intervalo de confiança de 95% da diferença entre a média populacional de saldos de conta corrente nos dois bancos filiais. Os dados amostrais exibem $n_1 = 28$, $\bar{x}_1 =$ US$ 1.025 e $s_1 =$ US$ 150 para a filial de Cherry Grove, e $n_1 = 22$, $\bar{x}_2 =$ US$ 910 e $s_2 =$ US$ 125 para a filial de Beechmont. O cálculo dos graus de liberdade para $t_{\alpha/2}$ é o seguinte:

$$gl = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{1}{n_1 - 1}\left(\frac{s_1^2}{n_1}\right)^2 + \frac{1}{n_2 - 1}\left(\frac{s_2^2}{n_2}\right)^2} = \frac{\left(\frac{150^2}{28} + \frac{125^2}{22}\right)^2}{\frac{1}{28 - 1}\left(\frac{150^2}{28}\right)^2 + \frac{1}{22 - 1}\left(\frac{125^2}{22}\right)^2} = 47{,}8$$

Arredondamos *para baixo* os graus de liberdade não-inteiros, para 47, para obtermos um valor de t ligeiramente maior e uma estimação por intervalo mais conservadora. Usando a tabela de distribuição t com 47 graus de liberdade, encontramos $t_{0,025} = 2{,}012$. Usando a Equação 10.6, desenvolvemos a estimação por intervalo de confiança de 95% da diferença entre as duas médias populacionais da seguinte maneira:

Esta sugestão o ajudará se você estiver usando a Equação 10.7 para calcular os graus de liberdade manualmente.

$$\bar{x}_1 - \bar{x}_2 \pm t_{0,025}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

$$1.025 - 910 \pm 2{,}012\sqrt{\frac{150^2}{28} + \frac{125^2}{22}}$$

$$115 \pm 78$$

A estimação por ponto da diferença entre a média populacional dos saldos de conta corrente nas duas filiais é US$ 115. A margem de erro é US$ 78 e a estimação por intervalo de confiança de 95% da diferença entre as duas médias populacionais é 115 – 78 = US$ 37 a 115 + 78 = US$ 193.

O cálculo dos graus de liberdade (Equação 10.7) é complicado se você o fizer manualmente, mas é facilmente implementado com um software. Note, porém, que as expressões s_1^2/n_1 e s_2^2/n_2 aparecem tanto na Equação 10.6 como na Equação 10.7. Esses valores precisam ser calculados somente uma vez para que se possa avaliar tanto a Equação 10.6 como a Equação 10.7.

Testes de Hipóteses sobre $\mu_1 - \mu_2$

Consideremos, agora, os testes de hipóteses a respeito da diferença entre as médias de duas populações quando os desvios padrão populacionais σ_1 e σ_2 são desconhecidos. Admitindo que D_0 denota a diferen-

ça hipotética entre μ_1 e μ_2, a Seção 10.1 mostrou que a estatística de teste usada para o caso em que σ_1 e σ_2 são conhecidos é a seguinte:

$$z = \frac{(\bar{x}_1 - \bar{x}_2) - D_0}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

A estatística de teste, z, segue a distribuição normal padrão.

Quando σ_1 e σ_2 são desconhecidos, usamos s_1 como um estimador de σ_1 e s_2 como um estimador de σ_2. Substituindo σ_1 e σ_2 por esses desvios padrão amostrais, obtemos a seguinte estatística de teste quando σ_1 e σ_2 são desconhecidos.

ESTATÍSTICA DE TESTE PARA TESTES DE HIPÓTESE SOBRE $\mu_1 - \mu_2$ QUANDO s_1 E s_2 SÃO DESCONHECIDOS

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - D_0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \qquad \textbf{(10.8)}$$

Os graus de liberdade para t são dados pela Equação 10.7.

Vamos demonstrar o uso dessa estatística de teste no seguinte exemplo de teste de hipóteses.

Considere um novo pacote de software desenvolvido para auxiliar analistas de sistemas a reduzir o tempo necessário para projetar, desenvolver e implementar sistemas de informação. Para avaliar os benefícios do novo pacote de software, uma amostra aleatória de 24 analistas de sistemas é selecionada. A cada analista são dadas as especificações de um sistema de informação hipotético. Então, 12 dos analistas são instruídos a produzir o sistema de informação utilizando a tecnologia atual. Os outros 12 analistas são treinados a usar o novo pacote de software e depois são instruídos a usá-lo para produzir o sistema de informação.

Esse estudo envolve duas populações: uma de analistas de sistemas que usam a tecnologia atual e uma de analistas de sistemas que usam o novo pacote de software. Em termos do tempo necessário para concluir o desenho do projeto de sistema de informação, as médias populacionais são as seguintes:

μ_1 = o tempo médio de conclusão do projeto para os analistas que usam a tecnologia atual.

μ_2 = o tempo médio de conclusão do projeto para os analistas que usam o novo pacote de software.

O pesquisador encarregado do projeto de avaliação do novo software espera demonstrar que o novo pacote de software apresentará uma média de tempo mais breve para a conclusão do projeto. Desse modo, o pesquisador está à procura de evidências que o levem a concluir que μ_2 é menor que μ_1; nesse caso, a diferença entre as duas médias populacionais, $\mu_1 - \mu_2$, será maior que zero. A hipótese de pesquisa $\mu_1 - \mu_2 > 0$ é declarada como a hipótese alternativa. Assim, o teste de hipóteses torna-se:

$$H_0: \mu_1 - \mu_2 \leq 0$$
$$H_a: \mu_1 - \mu_2 > 0$$

Usaremos $\alpha = 0{,}05$ como o nível de significância.

Suponha que os 24 analistas concluam o estudo com os resultados mostrados na Tabela 10.1. Usando a estatística de teste na Equação 10.8, obtemos:

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - D_0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{(325 - 286) - 0}{\sqrt{\frac{40^2}{12} + \frac{44^2}{12}}} = 2{,}27$$

Calculando os graus de liberdade com a Equação 10.7, obtemos,

$$gl = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{1}{n_1 - 1}\left(\frac{s_1^2}{n_1}\right)^2 + \frac{1}{n_2 - 1}\left(\frac{s_2^2}{n_2}\right)^2} = \frac{\left(\frac{40^2}{12} + \frac{44^2}{12}\right)^2}{\frac{1}{12 - 1}\left(\frac{40^2}{12}\right)^2 + \frac{1}{12 - 1}\left(\frac{44^2}{12}\right)^2} = 21{,}8$$

Arredondando, usaremos uma distribuição t com 21 graus de liberdade. Essa linha da tabela de distribuição t é a seguinte:

Área da Cauda Superior	0,20	0,10	0,05	0,025	0,01	0,005
Valor t (21 graus de liberdade)	0,859	1,323	1,721	2,080	2,518	2,831

t = 2,27

Usando a tabela de distribuição t, somente podemos determinar um intervalo para o valor *p*. O uso do computador é necessário para determinarmos o valor *p* exato.

Com um teste da cauda superior, o valor *p* é a área na cauda superior à direita de $t = 2{,}27$. Dos resultados anteriores, notamos que o valor *p* está entre 0,025 e 0,01. Desse modo, o valor *p* é menor que $\alpha = 0{,}05$, e H_0 é rejeitada. Os resultados amostrais possibilitam ao pesquisador concluir que $\mu_1 - \mu_2 > 0$, $\mu_1 > \mu_2$. Dessa forma, o estudo de pesquisa sustenta a conclusão de que o novo pacote de software oferece uma média populacional menor de tempo de conclusão.

Tabela 10.1 Dados sobre o tempo de conclusão e sumário estatístico do estudo dos testes do software

	Tecnologia Atual	Novo Software
	300	274
	280	220
	344	308
	385	336
	372	198
	360	300
	288	315
	321	258
	376	318
	290	310
	301	332
	283	263
Sumário Estatístico		
Tamanho da amostra	$n_1 = 12$	$n_2 = 12$
Média Amostral	$\bar{x}_1 = 325$ horas	$\bar{x}_2 = 286$ horas
Desvio padrão da amostra	$s_1 = 40$	$s_2 = 44$

ARQUIVO DA INTERNET SoftwareTest

O Minitab pode ser usado para analisar dados de teste de hipóteses sobre a diferença entre duas médias populacionais. A saída de dados (*output*) que compara a atual e a nova tecnologia de software é mostrada na Figura 10.2. A última linha da saída de dados apresenta $t = 2{,}27$ e o valor $p = 0{,}017$. Observe que o Minitab usou a Equação 10.7 para calcular 21 graus de liberdade para essa análise.

Conselho Prático

Os procedimentos de estimação por intervalo e de teste de hipóteses apresentados nesta seção são robustos e podem ser usados com tamanhos de amostra relativamente pequenos. Na maioria das aplicações, tamanhos de amostra iguais ou aproximadamente iguais, de forma que o tamanho de amostra total $n_1 + n_2$ seja, no mínimo, igual a 20, pode-se esperar que eles ofereçam resultados muito bons mesmo que as populações não sejam normais. Tamanhos de amostra maiores são recomendados se as distribuições das populações forem altamente assimétricas ou se tiverem pontos fora da curva. Tamanhos de amostra menores somente devem ser usados se o analista estiver convencido de que as distribuições das populações sejam, no mínimo, aproximadamente normais.

Sempre que possível, tamanhos de amostra iguais, $n_1 = n_2$, são recomendados.

Figura 10.2 Saída de dados do Minitab para o teste de hipóteses das tecnologias de software atual e nova

```
Two-sample T for Current vs New

          N      Mean     StDev    SE Mean
Current   12     325.0     40.0        12
New       12     286.0     44.0        13

Difference = mu Current - mu New
Estimate for difference: 39.0000
95% lower bound for difference = 9.4643
T-Test of difference = 0 (vs >):  T-Value = 2.27  P-Value = 0.017  DF = 21
```

NOTAS E COMENTÁRIOS

Outro critério usado para se fazer inferências sobre a diferença entre duas médias populacionais quando σ_1 e σ_2 são desconhecidos baseia-se na hipótese de que os dois desvios padrão populacionais são *iguais* ($\sigma_1 = \sigma_2 = \sigma$). Dessa hipótese, os dois desvios padrão amostrais são combinados para produzir a seguinte *variância amostral agrupada*:

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

A estatística de teste t torna-se:

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - D_0}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

e tem $n_1 + n_2 - 2$ graus de liberdade. Nesse ponto, o cálculo do valor p e a interpretação dos dados amostrais são idênticos aos procedimentos discutidos anteriormente nesta seção.

Uma dificuldade apresentada por esse procedimento é que a hipótese de que os dois desvios padrão são iguais geralmente é difícil de verificar. Desvios padrão populacionais não-iguais freqüentemente são encontrados. O uso do procedimento agrupado pode não fornecer resultados satisfatórios, especialmente se os tamanhos de amostra n_1 e n_2 forem muito diferentes.

O procedimento t que apresentamos nesta seção não requer a suposição de desvios padrão populacionais iguais e pode ser aplicado quer os desvios padrão populacionais sejam iguais ou não. É o procedimento mais geral e é recomendado para a maioria das aplicações.

Exercícios

Métodos

AUTOTESTE

9. Considere os seguintes resultados, correspondentes a amostras aleatórias independentes tomadas de duas populações.

Amostra 1	Amostra 2
$n_1 = 20$	$n_2 = 30$
$\bar{x}_1 = 22{,}5$	$\bar{x}_2 = 20{,}1$
$s_1 = 2{,}5$	$s_2 = 4{,}8$

a. Qual é a estimação por ponto da diferença entre as duas médias populacionais?
b. Qual é o grau de liberdade para a distribuição t?
c. Com 95% de confiança, qual é a margem de erro?
d. Qual é o intervalo de confiança para a diferença entre as duas médias populacionais?

10. Considere o seguinte teste de hipóteses.

$$H_0: \mu_1 - \mu_2 = 0$$
$$H_a: \mu_1 - \mu_2 \neq 0$$

Os resultados apresentados a seguir são de amostras independentes tomadas de duas populações:

Amostra 1	**Amostra 2**
$n_1 = 35$	$n_2 = 40$
$\bar{x}_1 = 13{,}6$	$\bar{x}_2 = 10{,}1$
$s_1 = 5{,}2$	$s_2 = 8{,}5$

a. Qual é o valor da estatística de teste?
b. Qual é o grau de liberdade da distribuição t?
c. Qual é o valor t
d. Com $\alpha = 0{,}05$, qual é a sua conclusão?

11. Considere os dados seguintes, correspondentes a amostras aleatórias independentes tomadas de duas populações normais:

Amostra 1	10	7	13	7	9	8
Amostra 2	8	7	8	4	6	9

a. Calcule a média das duas amostras.
b. Calcule os dois desvios padrão amostrais.
c. Qual é a estimação por ponto da diferença entre as duas médias populacionais?
d. Qual é a estimação por intervalo de confiança de 95% da diferença entre as duas médias populacionais?

Aplicações

12. O U.S. Department of Transportation divulga o número de milhas que os habitantes das 75 maiores regiões metropolitanas viajam de carro por dia. Suponha que, para uma amostra aleatória simples de 50 habitantes de Buffalo, a média seja de 22,5 milhas por dia e que o desvio padrão seja de 8,4 milhas por dia e que, para uma amostra aleatória simples independente de 40 habitantes de Boston, a média seja de 18,6 milhas por dia e que o desvio padrão seja de 7,4 milhas por dia.

a. Qual é a estimação por ponto da diferença entre o número médio de milhas que os habitantes de Buffalo viajam por dia e o número médio de milhas que os habitantes de Boston viajam por dia?
b. Qual é o intervalo de confiança de 95% da diferença entre as duas médias populacionais?

13. A FedEx e a United Parcel Service (UPS) são dois dos principais serviços de entrega de encomendas em termos de volume e receita (*The Wall Street Journal*, 27 de janeiro de 2004). De acordo com o Airports Council International, o Memphis International Airport (FedEx) e o Louisville International Airport (UPS) são dois dos maiores aeroportos de carga do mundo. As seguintes amostras aleatórias apresentam as toneladas de carga por dia manipuladas por esses dois aeroportos. Os dados estão expressos em milhares de toneladas.

Memphis

9,1	15,1	8,8	10,0	7,5	10,5
8,3	9,1	6,0	5,8	12,1	9,3

Louisville

4,7	5,0	4,2	3,3	5,5
2,2	4,1	2,6	3,4	7,0

a. Calcule a média amostral e o desvio padrão amostral correspondentes a cada aeroporto.
b. Qual é a estimação por ponto da diferença entre as duas médias populacionais? Interprete esse valor em termos do aeroporto que manipula o maior volume e de uma comparação da diferença de volume entre esses dois aeroportos.
c. Desenvolva um intervalo de confiança de 95% entre as médias populacionais diárias correspondentes aos dois aeroportos.

14. As áreas costeiras dos Estados Unidos, incluindo Cape Cod, as Outer Banks,[1] a Carolina do Norte e a Carolina do Sul e a Região Costeira do Golfo do México (*Gulf Coast*),[2] tiveram índices de crescimento populacional relativamente elevados durante a década de 1990. Foram coletados dados sobre os habitantes que vivem nas comunidades costeiras, bem como sobre os habitantes que vivem em áreas não-litorâneas de todas as regiões dos Estados Unidos (*USA Today*, 21 de julho de 2000). Suponha que os seguintes dados amostrais tenham sido obtidos sobre a idade das pessoas nas duas populações.

Áreas Costeiras	Áreas Não-Costeiras
$n_1 = 150$	$n_2 = 175$
$\bar{x}_1 = 39{,}3$ anos	$\bar{x}_2 = 35{,}4$ anos
$s_1 = 16{,}8$ anos	$s_2 = 15{,}2$ anos

Teste a hipótese de não haver nenhuma diferença entre as duas médias populacionais. Use $\alpha = 0{,}05$.
a. Formule as hipóteses nula e alternativa.
b. Qual é o valor da estatística de teste?
c. Qual é o valor *p*?
d. Qual é a sua conclusão?

15. Nos últimos anos, aumentaram as lesões nos jogadores da Major League Baseball. Em relação ao período de 1992 a 2001, a ampliação da liga fez que as inscrições à Major League Baseball aumentassem 15%. Entretanto, o número de jogadores que são colocados na lista de inativos em virtude das lesões aumentou 32% no mesmo período (*USA Today*, 8 de julho de 2002). Uma pergunta de pesquisa queria saber se os jogadores da Major League Baseball colocados na lista de inativos permaneciam nela durante um tempo mais longo em 2001 que os jogadores que eram colocados na lista de inativos há uma década.

a. Usando a média populacional do número de dias que um jogador permanece na lista de inativos, formule as hipóteses nula e alternativa que possam ser utilizadas para testar a pergunta da pesquisa.
b. Suponha que os seguintes dados sejam aplicáveis:

	Temporada de 2001	Temporada de 1992
Jogadores da Amostra	$n_1 = 45$	$n_2 = 38$
Média de Dias da Amostra	$\bar{x}_1 = 60$ dias	$\bar{x}_2 = 51$ dias
Desvio Padrão da Amostra	$s_1 = 18$ dias	$s_2 = 15$ dias

Qual é a estimação por ponto da diferença entre a média de dias da população que permanece na lista de inativos em 2001 em comparação com 1992? Qual é o aumento percentual no número de dias de permanência na lista de inativos?
c. Use $\alpha = 0{,}01$. Qual é a sua conclusão a respeito do número de dias de permanência na lista de inativos? Qual é o valor *p*?
d. Esses dados sugerem que a Major League Baseball deve preocupar-se com a situação?

ARQUIVO DA INTERNET SATVerbal

16. O College Board divulgou comparações sobre as pontuações no Scholastic Aptitude Test (SAT)[3] baseando-se no nível educacional mais elevado obtido pelos pais da pessoa que faz os exames. Uma das hipóteses de pesquisa era que os estudantes cujos pais haviam obtido um nível mais elevado de educação obteriam uma pontuação média mais elevada no SAT. Durante 2003, a média global dos exames orais do SAT foi 507 (*The World Almanac 2004*). As pontuações nos exames orais do SAT para amostras independentes de estudantes são apresentadas a seguir. A primeira amostra exibe pontuações nos exames orais do SAT correspondentes a estudantes cujos pais têm diplomas universitários com graus de bacharel. A segunda amostra exibe as pontuações nos exames orais do SAT de estudantes cujos pais têm diplomas do segundo grau, mas não têm diplomas universitários.

[1] NT: *Outer Banks* – Cadeia de ilhas arenosas longas e estreitas ao longo da costa da Carolina do Norte (Estados Unidos).

[2] NT: *Gulf Coast* – Estados do Golfo do México: Flórida, Alabama, Mississipi, Louisiana e Texas (Estados Unidos).

[3] NT: SAT, ou *Scholastic Aptitude Test* – É um exame usado pelas universidades como parte do processo de seleção de estudantes para admissão ao curso superior. Há sete seções: três de matemática, três orais e uma prática (experimental), que não recebe notas e é usada somente para pesquisa.

Pais do Estudante			
Com Diploma Universitário		Com Diploma do Segundo Grau	
485	487	442	492
534	533	580	478
650	526	479	425
554	410	486	485
550	515	528	390
572	578	524	535
497	448		
592	469		

a. Formule as hipóteses que podem ser usadas para determinar se os dados amostrais sustentam a hipótese de que os estudantes exibem uma pontuação média mais elevada nos exames orais do SAT se seus pais tiverem obtido um nível de educação mais elevado.
b. Qual é a estimação por ponto da diferença entre as médias das duas populações?
c. Calcule o valor p para o teste de hipóteses.
d. Com $\alpha = 0,05$, qual é a sua conclusão?

17. Periodicamente, os clientes da Merrill Lynch são solicitados a avaliar os consultores e os serviços financeiros dessa empresa (2000 Merrill Lynch Client Satisfaction Survey). Avaliações mais elevadas sobre a satisfação do cliente indicam um atendimento melhor, sendo 7 a classificação máxima para os serviços. Amostras independentes de avaliações do serviço prestado por dois consultores financeiros estão resumidas aqui. O consultor A tem dez anos de experiência, ao passo que o consultor B tem um ano de experiência. Use $\alpha = 0,05$ e teste para verificar se o consultor que tem mais experiência possui uma média de avaliação de atendimento populacional mais elevada.

Consultor A	**Consultor B**
$n_1 = 16$	$n_2 = 10$
$\bar{x}_1 = 6,82$	$\bar{x}_2 = 6,25$
$s_1 = 0,64$	$s_2 = 0,75$

a. Estabeleça as hipóteses nula e alternativa.
b. Calcule o valor da estatística de teste.
c. Qual é o valor p?
d. Qual é a sua conclusão?

18. As empresas de cursinhos universitários oferecem estudos dirigidos, aprendizagem em sala de aula e testes práticos, em um esforço para ajudar os estudantes a obterem melhor desempenho nos exames como o Scholastic Aptitude Test (SAT). As empresas de cursinhos universitários afirmam que seus cursos melhorarão o desempenho no SAT em uma média de 120 pontos (*The Wall Street Journal*, 23 de janeiro de 2003). Um pesquisador não tem tanta certeza a respeito dessa afirmação e acredita que 120 pontos podem ser uma afirmação exagerada no esforço para encorajar os estudantes a fazerem o cursinho. Em um estudo de avaliação do serviço prestado pelos cursinhos, o pesquisador coleta dados de pontuação no SAT de 35 estudantes que fizeram o cursinho e de 48 que não o fizeram.

a. Formule as hipóteses que podem ser usadas para testar a crença do pesquisador de que as pontuações obtidas no SAT podem ser menores que a média declarada de 120 pontos.
b. Use $\alpha = 0,05$ e os dados apresentados a seguir. Qual é a sua conclusão?

	Participantes do Cursinho	**Não-Participantes do Cursinho**
Média Amostral	1.058	983
Desvio Padrão da Amostra	90	105

c. Qual é a estimação por ponto da melhoria da média de pontuações no SAT proporcionada pelo cursinho universitário? Apresente uma estimação por intervalo de confiança de 95% da melhoria.
d. Qual conselho você daria ao pesquisador depois de ver o intervalo de confiança?

10.3 INFERÊNCIAS SOBRE A DIFERENÇA ENTRE AS MÉDIAS DE DUAS POPULAÇÕES: AMOSTRAS RELACIONADAS (OU DEPENDENTES)

Suponha que os empregados de uma empresa de manufatura possam usar dois diferentes métodos para executar uma tarefa de produção. Para maximizar o resultado da produção, a empresa quer identificar o método que apresenta a menor média populacional de tempo de conclusão. Digamos que μ_1 denote a média populacional do tempo de conclusão correspondente ao método de produção 1 e que μ_2 denote a média populacional do tempo de conclusão correspondente ao método de produção 2. Sem nenhuma indicação preliminar do método de produção preferido, iniciamos com a hipótese experimental de que os dois métodos têm a mesma média populacional de tempo de conclusão. Desse modo, a hipótese nula é H_0: $\mu_1 - \mu_2 = 0$. Se essa hipótese for rejeitada, podemos concluir que os tempos médios populacionais para a conclusão diferem. Nesse caso, o método que fornece o menor tempo médio de conclusão seria recomendado. As hipóteses nula e alternativa são escritas da seguinte maneira:

$$H_0: \mu_1 - \mu_2 = 0$$
$$H_a: \mu_1 - \mu_2 \neq 0$$

Ao escolher o procedimento de amostragem que usaremos para coletar os dados referentes ao tempo de produção e testar as hipóteses, consideramos duas alternativas de projeto. Uma se baseia em amostras independentes e a outra, em **amostras relacionadas**.

1. *Projeto de amostra independente*: Uma amostra aleatória simples de funcionários é selecionada e cada funcionário da amostra usa o método 1. Uma segunda amostra aleatória simples independente de funcionário é selecionada e cada funcionário dessa amostra usa o método 2. O teste da diferença entre as médias baseia-se nos procedimentos da Seção 10.2.
2. *Projeto de amostras relacionadas (ou combinadas)*: Uma amostra aleatória simples de funcionários é selecionada. Cada funcionário usa primeiramente um método e depois o outro. A ordem dos dois métodos é atribuída aleatoriamente aos funcionários, sendo que alguns executam primeiro o método 1 e os outros, o método 2. Cada funcionário produz um par de valores de dados, e um valor corresponde ao método 1 e outro, ao método 2.

No projeto de amostras relacionadas, os dois métodos de produção são testados sob condições idênticas (ou seja, com os mesmos funcionários); portanto, esse projeto acarreta um erro de amostragem menor que o projeto de amostras independentes. A razão básica para que isso ocorra é que em um projeto de amostras relacionadas as variações entre os trabalhadores são eliminadas porque são usadas as mesmas pessoas para ambos os métodos de produção.

Vamos demonstrar a análise de um projeto de amostras relacionadas presumindo que seja este o método utilizado para testar a diferença entre as médias populacionais dos dois métodos de produção. Uma amostra aleatória de seis funcionários é usada. Os dados sobre os tempos de conclusão da tarefa correspondentes aos seis funcionários são apresentados na Tabela 10.2. Note que cada funcionário fornece um par de valores de dados, sendo um para cada método de produção. Note também que a última coluna contém a diferença entre os tempos de conclusão d_1 correspondente a cada funcionário da amostra.

O elemento decisivo para a análise do projeto de amostras relacionadas é perceber que consideramos somente a coluna de diferenças. Portanto, temos seis valores de dados (0,6; –0,2; 0,5; 0,3; 0,0 e 0,6) que serão usados para analisar a diferença entre as médias populacionais dos dois métodos de produção.

ARQUIVO DA INTERNET
Matched

Tabela 10.2 Tempos de conclusão da tarefa correspondentes a um projeto de amostras relacionadas

Funcionário	Tempo de Conclusão para o Método (em minutos)	Tempo de Conclusão para o Método 2 (em minutos)	Diferença dos Tempos de Conclusão (d_i)
1	6,0	5,4	0,6
2	5,0	5,2	2,2
3	7,0	6,5	0,5
4	6,2	5,9	0,3
5	6,0	6,0	0,0
6	6,4	5,8	0,6

Admitamos que μ_d = a média dos valores de *diferença* para a população de funcionários. Com essa notação, as hipóteses nula e alternativa são reescritas da seguinte maneira:

$$H_0: \mu_d = 0$$
$$H_a: \mu_d \neq 0$$

Se H_0 for rejeitada, podemos concluir que a média populacional dos tempos de conclusão difere.

A notação d é um lembrete de que a amostra relacionada fornece dados de *diferença*. A média amostral e o desvio padrão amostral dos seis valores de diferença da Tabela 10.2 são os seguintes:

A não ser pelo uso da notação d, as fórmulas da média amostral e do desvio padrão amostral são as mesmas utilizadas anteriormente no texto.

$$\bar{d} = \frac{\Sigma d_i}{n} = \frac{1,8}{6} = 0,30$$

$$s_d = \sqrt{\frac{\Sigma(d_i - \bar{d})^2}{n-1}} = \sqrt{\frac{0,56}{0,5}} = 0,335$$

Com uma média amostral pequena de $n = 6$ trabalhadores, precisamos levantar a hipótese de que a população de diferenças tem uma distribuição normal. Essa hipótese é necessária a fim de podermos usar a distribuição t para os procedimentos de teste de hipóteses e de estimação por intervalo. Com base nessa hipótese, a seguinte estatística de teste tem uma distribuição t com $n - 1$ graus de liberdade:

Não é necessário levantar a hipótese de que a população tem uma distribuição normal se o tamanho da amostra for grande. Diretrizes sobre o tamanho de amostra para se usar a distribuição t foram apresentadas nos Capítulos 8 e 9.

$$t = \frac{\bar{d} - \mu_d}{s_d/\sqrt{n}} \tag{10.9}$$

Vamos usar a Equação 10.9 para testar as hipóteses H_0: $m_d = 0$ e H_a: $\mu_d \neq 0$, usando a = 0,05. Substituindo os resultados amostrais $\bar{d} = 0,30$, $s_d = 0,335$ e $n = 6$ na Equação 10.9, calculamos o valor da estatística de teste.

$$t = \frac{\bar{d} - \mu_d}{s_d/\sqrt{n}} = \frac{0,30 - 0}{0,335/\sqrt{6}} = 2,20$$

Vamos calcular o valor p para esse teste bicaudal. Uma vez que $t = 2,20 > 0$, a estatística de teste está na cauda superior da distribuição t. Com $t = 2,20$, a área na cauda superior à direita da estatística de teste pode ser encontrada usando-se a tabela de distribuição t com graus de liberdade $= n - 1 = 6 - 1 = 5$.

A informação contida na linha de 5 graus de liberdade da tabela de distribuição t é a seguinte:

Área da Cauda Superior	0,20	0,10	0,05	0,025	0,01	0,005
Valor t (5 graus de liberdade)	0,920	1,476	2,015	2,571	3,365	4,032

t = 2,20

Tão logo os dados de diferença são calculados, o procedimento de distribuição t para amostras relacionadas é idêntico aos procedimentos de estimação de uma população e de teste de hipóteses descritos nos Capítulos 8 e 9.

Desse modo, notamos que a área na cauda superior está entre 0,05 e 0,025. Uma vez que este teste é um teste bicaudal, duplicamos esses valores e concluímos que o valor p está entre 0,10 e 0,05. Esse valor p é maior que $\alpha = 0,05$. Assim, a hipótese nula H_0: $\mu_d = 0$ não é rejeitada. Usando o Minitab e os dados da Tabela 10.2, encontramos o valor $p = 0,080$.

Além disso, podemos obter uma estimação por intervalo da diferença entre as duas médias populacionais usando a metodologia das populações simples apresentada no Capítulo 8. Com 95% de confiança, o cálculo é o seguinte:

$$\bar{d} \pm t_{0,025}\frac{s_d}{\sqrt{n}}$$

$$0,3 \pm 2,571\left(\frac{0,335}{\sqrt{6}}\right)$$

$$0,3 \pm 0,35$$

Portanto, a margem de erro é 0,35, e o intervalo de confiança de 95% para a diferença entre as médias populacionais dos dois métodos de produção é de – 0,05 minutos a 0,65 minutos.

NOTAS E COMENTÁRIOS

1. No exemplo apresentado nesta seção, funcionários executavam a tarefa de produção utilizando primeiramente um dos métodos e depois o outro. Esse exemplo ilustra um projeto de amostras relacionadas no qual cada elemento amostrado (funcionário) produz um par de valores de dados. Também é possível usar elementos diferentes, porém, "similares", para produzir um par de valores de dados. Por exemplo, poderia haver uma correspondência de um trabalhador situado em um lugar com um trabalhador similar situado em outro lugar (sendo a correspondência baseada em idade, educação, sexo, experiência profissional etc.). Os pares de trabalhadores produziriam os dados da diferença que poderiam ser usados na análise de amostras relacionadas.
2. Um procedimento de amostras relacionadas (ou pendentes) para inferências sobre duas médias populacionais geralmente produz melhor precisão que o critério de amostras independentes; portanto, ele é o projeto recomendado. Entretanto, em algumas aplicações, a correspondência não pode ser obtida ou, talvez, o tempo e o custo associados com a correspondência sejam excessivos. Nesses casos, o projeto de amostras independentes deve ser usado.

Exercícios

Métodos

AUTOTESTE

19. Considere o seguinte teste de hipóteses:

$$H_0: \mu_d \leq 0$$
$$H_a: \mu_d > 0$$

Os dados a seguir são de amostras relacionadas tomadas de duas populações.

Elemento	População 1	População 2
1	21	20
2	28	26
3	18	18
4	20	20
5	26	24

a. Calcule o valor da diferença correspondente a cada elemento.
b. Calcule $\bar{d}$.
c. Calcule o desvio padrão s_d.
d. Realize um teste de hipóteses usando $\alpha = 0{,}05$. Qual é a sua conclusão?

20. Os dados a seguir são de amostras relacionadas que foram tomadas de duas populações.

Elemento	População 1	População 2
1	11	8
2	7	8
3	9	6
4	12	7
5	13	10
6	15	15
7	15	14

a. Calcule o valor da diferença correspondente a cada elemento.
b. Calcule $\bar{d}$.
c. Calcule o desvio padrão s_d.
d. Qual é a estimação por ponto da diferença entre as duas médias populacionais?
e. Forneça um intervalo de confiança de 95% da diferença entre as duas médias populacionais.

Aplicações

21. Uma firma de pesquisa de mercado usou uma amostra de indivíduos para avaliar o potencial de compra de determinado produto antes e depois de as pessoas virem um novo comercial de televisão a respeito do produto. As avaliações do potencial de compra basearam-se em uma escala de 0 a 10, e os valores mais altos indicavam maior potencial de compra. A hipótese nula declarava que a avaliação média "depois" seria menor ou igual à avaliação média "antes". A rejeição dessa hipótese demonstraria que o comercial melhorou a avaliação do potencial médio de compra. Use $\alpha = 0{,}05$ e os dados apresentados a seguir para testar a hipótese e comentar o valor do comercial.

	Avaliação de Compra			Avaliação de Compra	
Indivíduo	Depois	Antes	Indivíduo	Depois	Antes
1	6	5	5	3	5
2	6	4	6	9	8
3	7	7	7	7	5
4	4	3	8	6	6

22. Uma amostra de dez chamadas telefônicas internacionais forneceu o preço das tarifas por minuto da Sprint e da WorldCom para chamadas feitas dos Estados Unidos (*World Traveler*, julho de 2000).

País	Sprint	WorldCom
Austrália	0,46	0,26
Bélgica	0,69	0,40
Brasil	0,92	0,53
Colômbia	0,55	0,53
Dinamarca	0,50	0,26
França	0,46	0,26
Alemanha	0,46	0,26
Hong Kong	0,92	0,40
Japão	0,69	0,40
Reino Unido	0,46	0,26

Forneça uma estimação por intervalo de confiança de 95% da diferença entre as duas médias populacionais.

23. A Consumer Spending Survey (Pesquisa de Gastos de Consumo) do Bank of America coletou dados sobre os gastos anuais com cartão de crédito em sete diferentes categorias: transporte, produtos de mercearia, restaurantes, despesas domésticas, mobiliário doméstico, vestuário e entretenimento (*U.S. Airways Attaché*, dezembro de 2003). Usando dados de uma amostra de 42 contas de cartão de crédito, suponha que cada conta tenha sido usada para identificar os gastos anuais com cartão de crédito em produtos de mercearia (população 1) e os gastos anuais com cartão de crédito em restaurantes (população 2). Usando os dados de diferença, a diferença média amostral foi de $\bar{d}$ = US$ 850, e o desvio padrão amostral, de s_d = US$ 1.123.
    a. Formule as hipóteses nula e alternativa para testar se não há diferença entre a média populacional de gastos com cartão de crédito em produtos de mercearia e a média populacional de gastos com cartão de crédito em restaurantes.
    b. Use o nível de significância de 0,05. Você pode concluir que as médias populacionais diferem? Qual é o valor *p*?
    c. Qual categoria, a de produtos de mercearia ou a de restaurantes, tem uma média populacional mais elevada no que diz respeito aos gastos anuais com cartão de crédito? Qual é a estimação por ponto do intervalo de confiança de 95% da diferença entre as médias populacionais?

24. Os preços por galão (3,78 litros) de gasolina para carros de aluguel foram amostrados em oito grandes aeroportos. Os dados relativos às empresas de carros de aluguel Hertz e National são apresentados a seguir (*USA Today*, 4 de abril de 2000).

Aeroporto	Hertz	National
Boston Logan	1,55	1,56
Chicago O'Hare	1,62	1,59
Los Angeles	1,72	1,78
Miami	1,65	1,49

Aeroporto	Hertz	National
Nova York (JFK)	1,72	1,51
Nova York (LaGuardia)	1,67	1,50
Orange County, CA	1,68	1,77
Washington (Dulles)	1.52	1,41

Use $\alpha = 0,05$ para testar a hipótese de que não há nenhuma diferença entre os preços médios populacionais por galão em relação às duas empresas.

25. Nos últimos anos, uma sucessão crescente de opções de entretenimento compete pela atenção dos clientes. Em 2004, televisão a cabo e o rádio suplantaram a televisão convencional, as gravações musicais e os noticiários diários e se transformaram nas duas mídias de entretenimento mais utilizadas (*The Wall Street Journal*, 26 de janeiro de 2004). Pesquisadores usaram uma amostra de 15 indivíduos e coletaram dados sobre as horas por semana que eles passam a assistir à TV a cabo e as horas por semana que ouvem rádio.

ARQUIVO DA INTERNET
TVRadio

Indivíduo	Televisão	Rádio	Indivíduo	Televisão	Rádio
1	22	25	9	21	21
2	8	10	10	23	23
3	25	29	11	14	15
4	22	19	12	14	18
5	12	13	13	14	17
6	26	28	14	16	15
7	22	23	15	24	23
8	19	21			

a. Use o nível de significância 0,05 e teste se há alguma diferença entre a média populacional de uso da TV a cabo e do rádio. Qual é o valor p?
b. Qual é o número médio amostral de horas por semana que eles assistem à TV a cabo? Qual é o número médio amostral de horas por semana que eles ouvem rádio? Qual meio de comunicação tem o maior uso?

26. A StreetInsider.com divulgou dados referentes aos rendimentos por ação das maiores empresas em 2002 (12 de fevereiro de 2003). Antes de 2002, analistas financeiros fizeram previsões dos rendimentos por ação em 2002 para essas mesmas empresas (*Barron's*, 10 de setembro de 2001). Use os dados a seguir para comentar as diferenças entre os rendimentos por ação reais e os rendimentos por ação previstos.

ARQUIVO DA INTERNET
Earnings

Empresa	Reais	Previstos
AT&T	1,29	0,38
American Express	2,01	2,31
Citigroup	2,59	3,43
Coca-Cola	1,60	1,78
DuPont	1,84	2,18
Exxon-Mobil	2,72	2,19
General Electric	1,51	1,71
Johnson & Johnson	2,28	2,18
McDonald's	0,77	1,55
Wal-Mart	1,81	1,74

a. Use $\alpha = 0,05$ e teste se há alguma diferença entre a média populacional real e a média populacional prevista dos rendimentos por ação. Qual é o valor p? Qual é a sua conclusão?
b. Qual é a estimação por ponto da diferença entre as duas médias? Os analistas tenderam a subestimar ou a superestimar os rendimentos?
c. Com 95% de confiança, qual é a margem de erro para a estimativa do item (b)? O que você recomendaria com base nessa informação?

10.4 INTRODUÇÃO À ANÁLISE DE VARIÂNCIA

Até agora, enfatizamos os procedimentos estatísticos utilizados para comparar duas médias populacionais. Nesta seção, apresentamos a **análise de variância (ANOVA)** e mostramos como ela pode ser usada para

testar as hipóteses de que três ou mais populações são iguais. Iniciamos a discussão considerando um problema enfrentado pela National Computer Products, Inc.

Se H_0 for rejeitada, não poderemos concluir que todas as médias populacionais sejam diferentes. Rejeitar H_0 significa que pelo menos duas médias populacionais têm valores diferentes.

A National Computer Products, Inc. (NCP) produz impressoras e máquinas de fax em suas fábricas localizadas em Atlanta, Dallas e Seattle. Para medir quanto os empregados dessas fábricas sabem sobre gerenciamento da qualidade total, uma amostra aleatória de seis empregados de cada fábrica foi selecionada e seus integrantes foram submetidos a um exame de seus conhecimentos sobre a qualidade. As notas de exame obtidas por esses 18 empregados se encontram na Tabela 10.3. As médias amostrais, as variâncias amostrais e os desvios padrão amostrais de cada grupo também são apresentados. Os gerentes querem usar esses dados para testar a hipótese de que a média das notas de exame é a mesma para todas as três fábricas.

Definiremos a população 1 como todos os empregados da fábrica em Atlanta, a população 2 como todos os empregados da fábrica em Dallas e a população 3 como todos os empregados da fábrica em Seattle. Admitamos que:

μ_1 = média das notas de exame da população 1
μ_2 = média das notas de exame da população 2
μ_3 = média das notas de exame da população 3

Embora jamais saibamos os valores reais de μ_1, μ_2 e μ_3, queremos usar os resultados amostrais para testar as seguintes hipóteses.

H_0: $\mu_1 = \mu_2 = \mu_3$
H_a: Nem todas as médias populacionais são iguais

Conforme demonstraremos em breve, a análise de variância é um procedimento estatístico que pode ser usado para determinar se as diferenças observadas nas três médias amostrais são suficientemente grandes para rejeitarmos H_0.

Na introdução deste capítulo, afirmamos que a análise de variância pode ser usada para analisar dados obtidos tanto de um estudo observacional como de um estudo experimental. Para contarmos com uma nomenclatura comum para discutir o uso da análise de variância em ambos os tipos de estudo, precisamos introduzir os conceitos de variável de resposta, fator e tratamento.

As duas variáveis do exemplo da NCP são: a localização das fábricas e as notas obtidas no exame de conhecimento sobre qualidade. Uma vez que o objetivo é determinar se a média das notas de exame é a mesma para as fábricas localizadas em Atlanta, Dallas e Seattle, as notas de exame são chamadas variável dependente ou *variável de resposta* e o local da fábrica como a variável independente ou *fator*. Em geral, os valores de um fator selecionado para serem submetidos a uma investigação denominam-se níveis do fator ou *tratamentos*. Desse modo, no exemplo da NCP, os três tratamentos são Atlanta, Dallas e Seattle. Esses três tratamentos definem as populações de interesse no exemplo da NCP. Para cada tratamento, ou população, a variável de resposta é a nota obtida no exame.

Se os tamanhos de amostra forem iguais, a análise de variância não terá sensibilidade suficiente para detectar afastamentos da hipótese de que as populações estão normalmente distribuídas.

Tabela 10.3 Notas de exame dos 18 empregados

Observação	Fábrica 1 Atlanta	Fábrica 2 Dallas	Fábrica 3 Seattle
1	85	71	59
2	75	75	64
3	82	73	62
4	76	74	69
5	71	69	75
6	85	82	67
Média amostral	79	74	66
Variação Amostral	34	20	32
Desvio padrão amostral	5,83	4,47	5,66

ARQUIVO DA INTERNET
NCP

Hipóteses sobre a Análise de Variância

Três hipóteses são necessárias para a análise de variância.

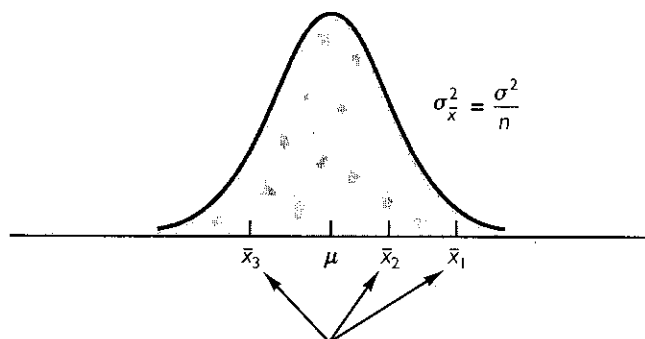
1. **Para cada população, a variável de resposta está normalmente distribuída.** Implicação: No exemplo da NCP, as notas obtidas no exame (variável de resposta) devem estar normalmente distribuídas em cada fábrica.
2. **A variância da variável de resposta, denotada por σ^2, é idêntica para todas as populações.** Implicação: No exemplo da NCP, a variância das notas obtidas no exame deve ser idêntica para todas as três fábricas.
3. **As observações devem ser independentes.** Implicação: No exemplo da NCP, a nota que cada empregado obteve no exame deve ser independente daquela obtida por qualquer outro empregado.

Visão Conceitual

Se as médias correspondentes às três populações fossem iguais, esperaríamos que as três médias amostrais estivessem bem próximas entre si. Realmente, quanto mais próximas as três médias amostrais estiverem entre si, mais evidências teremos para a conclusão de que as médias populacionais são iguais. Alternativamente, quanto mais diferirem as médias amostrais, mais evidências teremos para a conclusão de que as médias populacionais não são iguais. Em outras palavras, se a variabilidade entre as médias amostrais for "pequena", ela exibirá evidências favoráveis a H_0; se a variabilidade entre as médias amostrais for "grande", ela exibirá evidências favoráveis a H_a.

Se a hipótese nula, H_0: $\mu_1 = \mu_2 = \mu_3$, for verdadeira, poderemos usar a variabilidade entre as médias amostrais para desenvolver uma estimativa de σ^2. Primeiramente, observe que se as hipóteses referentes à análise de variância forem satisfeitas, cada amostra será proveniente da mesma distribuição normal com média μ e variância σ^2. Lembre-se do Capítulo 7 que a distribuição amostral da média $\bar{x}$ da amostra correspondente a uma amostra aleatória simples de tamanho n extraída de uma população normal, estará normalmente distribuída e possui uma média μ com uma variância σ^2/n. A Figura 10.3 ilustra esse tipo de distribuição amostral.

Desse modo, se a hipótese nula for verdadeira, podemos imaginar cada uma das três médias amostrais $\bar{x}_1 = 79$, $\bar{x}_2 = 74$ $\bar{x}_3 = 66$, apresentadas na Tabela 10.3, como valores extraídos aleatoriamente da distribuição amostral exibida na Figura 10.3. Nesse caso, a média e a variância dos três valores $\bar{x}$ podem ser usadas para estimar a média e a variância da distribuição amostral. Quando os tamanhos de amostra são iguais, como no exemplo da NCP, a melhor estimativa da média da distribuição amostral de $\bar{x}_1$ é a média, ou valor médio, das médias amostrais. Assim, no exemplo da NCP, uma estimativa da média da distribuição amostral de $\bar{x}$ é $(79 + 74 + 66)/3 = 73$. Referimo-nos a essa estimativa como *média global da amostra*. Uma estimativa da variância da distribuição amostral de $\bar{x}$, $\sigma_{\bar{x}}^2$, é fornecida pela variância das três médias amostrais:

Figura 10.3 Distribuição amostral de x, dado que H_0 seja verdadeira

$$s_{\bar{x}}^2 = \frac{(79 - 73)^2 + (74 - 73)^2 + (66 - 73)^2}{3 - 1} = \frac{86}{2} = 43$$

Uma vez que $\sigma_{\bar{x}}^2 = \sigma^2/n$, a resolução de σ^2 fornece

$$\sigma^2 = n\sigma_{\bar{x}}^2$$

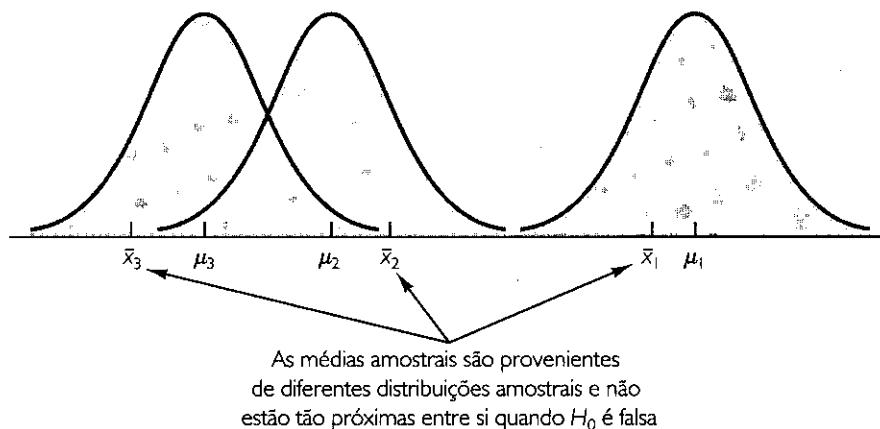
Portanto,

$$\text{Estimativa de } \sigma^2 = n \text{ (Estimativa de } \sigma_{\bar{x}}^2) = ns_{\bar{x}}^2 = 6(43) = 258$$

O resultado, $ns_{\bar{x}}^2 = 258$, denomina-se estimativa de σ^2 *entre tratamentos.*

A estimativa de σ^2 entre tratamentos baseia-se na suposição de que a hipótese nula é verdadeira. Nesse caso, cada amostra é proveniente da mesma população e há somente uma distribuição amostral de $\bar{x}$. Para ilustrar o que acontece quando H_0 é falsa, suponha que todas as médias populacionais difiram. Note que, desde que as três amostras sejam provenientes de populações normais com diferentes médias, elas resultarão em três diferentes distribuições amostrais. A Figura 10.4 mostra que, nesse caso, as médias amostrais não estão tão próximas como estavam quando H_0 era verdadeira. Dessa forma, $s_{\bar{x}}^2$ será maior, fazendo que a estimativa de σ^2 entre tratamentos seja maior. Em geral, quando as médias populacionais não são iguais, a estimativa entre tratamentos superestima a variância populacional σ^2.

A variância que ocorre dentro de cada uma das amostras também tem um efeito sobre a conclusão a que chegamos ao realizar a análise de variância. Quando uma amostra aleatória simples é selecionada de cada população, cada uma das variâncias amostrais fornece uma estimativa sem viés de σ^2. Portanto, podemos combinar ou agrupar as estimativas individuais de σ^2 em uma estimativa global. A estimativa global de σ^2 obtida dessa maneira é chamada *estimativa agrupada* ou *estimativa de σ^2 dentro dos tratamentos.* Uma vez que cada variância amostral fornece uma estimativa de σ^2 baseada somente na variação existente dentro de cada amostra, a estimativa de σ^2 dentro dos tratamentos não é afetada pelo fato de as médias populacionais serem ou não serem iguais.

Figura 10.4 Distribuições amostrais de $\bar{x}$, dado que H_0 seja falsa

Quando os tamanhos das amostras são iguais, a estimativa de σ^2 dentro dos tratamentos pode ser obtida calculando-se a média das variâncias amostrais individuais. Para o exemplo da NCP, obtemos

$$\text{Estimativa de } \sigma^2 \text{ dentro dos tratamentos} = \frac{34 + 20 + 32}{3} = \frac{86}{3} = 28{,}67$$

No exemplo da NCP, a estimativa de σ^2 entre tratamentos (258) é muito maior que a estimativa de σ^2 dentro dos tratamentos (28,67). Realmente, a razão dessas duas estimativas é 258/28,67 = 9,00. Lembre-se, porém, de que a abordagem entre tratamentos produz uma boa estimativa de σ^2 somente se a hipótese

nula for verdadeira; se a hipótese nula for falsa, a abordagem entre tratamentos superestimará σ^2. O critério dentro do tratamento fornece uma boa estimativa de σ^2 em qualquer um dos casos. Desse modo, se a hipótese nula for verdadeira, as duas estimativas serão similares e suas razões serão próximas de 1. Se a hipótese nula for falsa, a estimativa entre tratamentos será maior que a estimativa dentro dos tratamentos, e a razão entre elas será grande. Na próxima seção, mostraremos qual deve ser o tamanho dessa razão para rejeitarmos H_0.

Em suma, a lógica que há por trás da análise de variância (ANOVA) baseia-se no desenvolvimento de duas estimativas independentes da variância populacional σ^2 comum. Uma estimativa de σ^2 baseia-se na variabilidade existente entre as próprias médias amostrais, e a outra estimativa de σ^2 baseia-se na variabilidade dos dados existentes dentro de cada amostra. Ao comparar essas duas estimativas de σ^2, seremos capazes de determinar se as médias populacionais são iguais.

NOTAS E COMENTÁRIOS

Nas Seções 10.1 e 10.2, apresentamos métodos estatísticos para testar as hipóteses de que duas médias populacionais são iguais. A ANOVA também pode ser usada para testar as hipóteses de que duas médias populacionais são iguais. Na prática, entretanto, a análise de variância geralmente não é usada enquanto não se lida com três ou mais médias populacionais.

10.5 ANÁLISE DE VARIÂNCIA: COMO TESTAR A IGUALDADE DE *k* MÉDIAS DA POPULAÇÃO

A análise de variância pode ser usada para testar a igualdade de k médias populacionais. A forma geral das hipóteses testadas é:

$$H_0\colon \mu_1 = \mu_2 = \ldots = \mu_k$$
$$H_a\colon \text{Nem todas as médias populacionais são iguais}$$

em que

$$\mu_j = \text{média da } j\text{-ésima população}$$

Supomos que a amostra aleatória simples de tamanho n_j tenha sido selecionada de cada uma das k populações ou tratamentos. Em relação aos dados amostrais resultantes, admitimos que:

x_{ij} = valor da observação i para o tratamento j
n_j = número de observações para o tratamento j
$\bar{x}_j$ = média amostral para o tratamento j
s_j^2 = variância amostral para o tratamento j
s_j = desvio padrão amostral para o tratamento j

As fórmulas correspondentes à média amostral e à variância amostral para o tratamento j são as seguintes:

$$\bar{x}_j = \frac{\sum_{i=1}^{n_j} x_{ij}}{n_j} \quad \textbf{(10.10)}$$

$$s_j^2 = \frac{\sum_{i=1}^{n_j} (x_{ij} - \bar{x}_j)^2}{n_j - 1} \quad \textbf{(10.11)}$$

A média global da amostra, denotada por $\bar{\bar{x}}$, é a soma de todas as observações dividida pelo número total de observações. Ou seja,

$$\bar{\bar{x}} = \frac{\sum_{j=1}^{k}\sum_{i=1}^{n_j} x_{ij}}{n_T} \quad (10.12)$$

em que

$$n_T = n_1 + n_2 + \cdots + n_k \quad (10.13)$$

Se o tamanho de cada amostra for n, $n_T = kn$. Nesse caso, a Equação (10.12) se reduz a:

$$\bar{\bar{x}} = \frac{\sum_{j=1}^{k}\sum_{i=1}^{n_j} x_{ij}}{kn} = \frac{\sum_{j=1}^{k}\sum_{i=1}^{n_j} x_{ij}/n}{k} = \frac{\sum_{j=1}^{k} \bar{x}_j}{k} \quad (10.14)$$

Em outras palavras, quando quer que os tamanhos de amostra sejam iguais, a média global da amostra é simplesmente o valor médio das k médias amostrais.

Uma vez que cada amostra no exemplo da NCP consiste em $n = 6$ observações, a média global de amostra pode ser calculada usando-se a Equação (10.14). Para os dados da Tabela 10.3, obtivemos o seguinte resultado:

$$\bar{\bar{x}} = \frac{79 + 74 + 66}{3} = 73$$

Se a hipótese nula for verdadeira ($\mu_1 = \mu_2 = \mu_3 = \mu$). A média global da amostra igual a 73 será a melhor estimativa da média populacional μ.

Estimativa da Variância Populacional entre Tratamentos

Na seção anterior, introduzimos o conceito de estimativa de σ^2 entre tratamentos e mostramos como calculá-la quando os tamanhos de amostra são iguais. Essa estimativa de σ^2 é chamada *quadrado da média em razão do tratamento* e é denotada por MSTR (*mean square due to treatments*). A fórmula geral para calcular a MSTR é:

$$\text{MSTR} = \frac{\sum_{j=1}^{k} n_j(\bar{x}_j - \bar{\bar{x}})^2}{k-1} \quad (10.15)$$

O numerador da Equação (10.15) é chamado *soma dos quadrados dos tratamentos* e é denotado por SSTR (*sum of squares due to treatments*). O denominador, $k - 1$, representa os graus de liberdade associados à SSTR. Portanto, o quadrado médio dos tratamentos pode ser calculado pela seguinte fórmula.

QUADRADO MÉDIO DOS TRATAMENTOS

$$\text{MSTR} = \frac{\text{SSTR}}{k-1} \quad (10.16)$$

em que

$$\text{SSTR} = \sum_{j=1}^{k} n_j(\bar{x}_j - \bar{\bar{x}})^2 \quad (10.17)$$

Se H_0 for verdadeira, a MSTR produzirá uma estimativa sem viés de σ^2. Entretanto, se as médias de k populações não forem iguais, a MSTR não será uma estimativa sem viés de σ^2; realmente, nesse caso, a MSTR deve superestimar σ^2.

Em relação aos dados da NCP apresentados na Tabela 10.3, obtemos os seguintes resultados:

$$\text{SSTR} = \sum_{j=1}^{k} n_j(\bar{x}_j - \bar{\bar{x}})^2 = 6(79 - 73)^2 + 6(74 - 73)^2 + 6(66 - 73)^2 = 516$$

$$\text{MSTR} = \frac{\text{SSTR}}{k - 1} = \frac{516}{2} = 258$$

Estimativa da Variância Populacional dentro de Tratamentos

Anteriormente, introduzimos o conceito de estimativa de σ^2 dentro de tratamentos e mostramos como calculá-la quando os tamanhos de amostra fossem iguais. Essa estimativa de σ^2 é chamada *quadrado médio dos erros*, e é denotada por MSE (*mean square due to error*). A fórmula geral para calcular o MSE é:

$$\text{MSE} = \frac{\sum_{j=1}^{k}(n_j - 1)s_j^2}{n_T - k} \qquad \textbf{(10.18)}$$

O numerador da Equação 10.18 é denominado *soma dos quadrados dos erros* e é denotado por SSE (*sum of squares due to error*). O denominador de MSE, $n_T - k$, é o grau de liberdade associado à SSE. Portanto, a fórmula para calcular MSE também pode ser definida da seguinte forma:

QUADRADO MÉDIO DOS ERROS

$$\text{MSE} = \frac{\text{SSE}}{n_T - k} \qquad \textbf{(10.19)}$$

em que

$$\text{SSE} = \sum_{j=1}^{k}(n_j - 1)s_j^2 \qquad \textbf{(10.20)}$$

Observe que o MSE baseia-se na variação dentro de cada um dos tratamentos; ele não é influenciado pelo fato de a hipótese nula ser ou não ser verdadeira. Desse modo, o MSE sempre produz uma estimativa sem viés de σ^2.

Em relação aos dados da NCP apresentados na Tabela 10.3, obtemos os seguintes resultados:

$$\text{SSE} = \sum_{j=1}^{k}(n_j - 1)s_j^2 = (6 - 1)34 + (6 - 1)20 + (6 - 1)32 = 430$$

$$\text{MSE} = \frac{\text{SSE}}{n_T - k} = \frac{430}{18 - 3} = \frac{430}{15} = 28{,}67$$

Comparando as Estimativas de Variância: o Teste *F*

Se a hipótese nula for verdadeira, o MSTR e o MSE produzem duas estimativas independentes da variância populacional σ^2. Quando a hipótese nula é verdadeira e as pressuposições ANOVA são válidas, a distribuição amostral da razão MSTR/MSE tem uma **distribuição *F*** com $k - 1$ graus de liberdade no numerador e $n_T - 1$ graus de liberdade no denominador. A forma geral dessa distribuição F é mostrada na Figura 10.5. Se a hipótese nula for verdadeira, o valor de MSTR/MSE parecerá que é proveniente dessa distribuição. Entretanto, se a hipótese nula for falsa, o valor de MSTR/MSE sofrerá uma inflação, porque um MSTR grande produz uma estimativa em excesso de σ^2. Os valores de MSTR/MSE que levam à rejeição da hipótese nula estarão na cauda superior da distribuição mostrada na Figura 10.5.

Com a decisão de rejeitar a hipótese nula H_0 baseando-se na razão MSTR/MSE, essa razão torna-se a estatística de teste do teste de hipóteses sobre a igualdade de k médias populacionais. A estatística de teste é a seguinte:

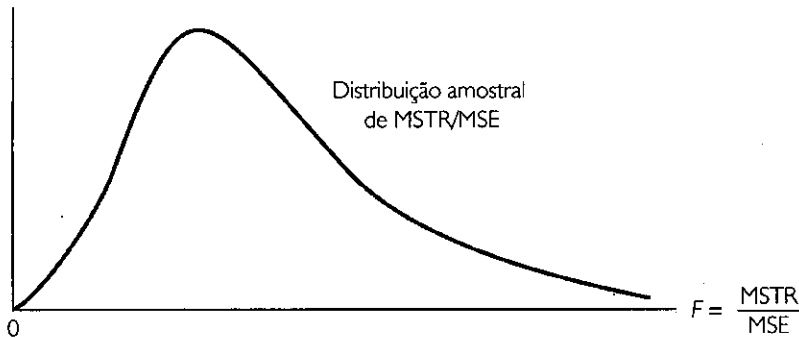
ESTATÍSTICA DE TESTE DA IGUALDADE DE k MÉDIAS POPULACIONAIS

$$F = \frac{\text{MSTR}}{\text{MSE}} \qquad \textbf{(10.21)}$$

A distribuição F tem $k - 1$ graus de liberdade no numerador e $n_T - k$ graus de liberdade no denominador.

Retornemos ao exemplo da National Computer Products e usemos um nível de significância $\alpha = 0{,}05$ para realizar o teste de hipóteses. As hipóteses nula e alternativa são redefinidas da seguinte maneira:

H_0: $\mu_1 = \mu_2 = \mu_3$
H_a: Nem todas as médias populacionais são iguais

Figura 10.5 Distribuição F: a distribuição amostral de MSTR/MSE

Com MSTR = 258 e MSE = 28,67 calculados anteriormente, o valor da estatística de teste é

$$F = \frac{\text{MSTR}}{\text{MSE}} = \frac{258}{28{,}67} = 9$$

Os graus de liberdade do numerador são $k - 1 = 3 - 1 = 2$, e os graus de liberdade do denominador são $n_T - k = 18 - 3 = 15$. Uma vez que rejeitamos a hipótese nula para valores grandes da estatística de teste, calcularemos o valor p como a área da cauda superior da distribuição F à direita da estatística de teste $F = 9$. A regra de rejeição de H_0 dos testes de hipótese habituais se o valor $p \leq \alpha$ aplica-se nesse caso.

A Tabela 10.4 apresenta uma parte da tabela de distribuição F que será útil nesse exemplo. Usando 2 graus de liberdade no numerador e 15 graus de liberdade no denominador, essa tabela exibe as seguintes áreas na cauda superior:

Área da Cauda Superior	0,10	0,05	0,025	0,01
Valor F ($gl_1 = 2$, $gl_2 = 15$)	2,70	3,68	4,77	6,36

$F = 9$

Já que $F = 9$ é maior que 6,36, a área da cauda superior em $F = 9$ é menor que 0,01. Desse modo, o valor p é menor que 0,01. Com o valor $p \leq \alpha = 0{,}05$, H_0 é rejeitada. O teste fornece suficientes evidências para concluirmos que as médias das três populações não são iguais. Em outras palavras, a análise de variância sustenta a conclusão de que a média populacional das notas de exame nas três fábricas da NCP não é igual.

Visto que a tabela F somente fornece valores para áreas da cauda superior correspondentes a 0,10, 0,05, 0,025 e 0,01, não podemos determinar o valor p exato diretamente da tabela. O Minitab ou o Excel fornecem o valor p como parte da saída de dados padrão ANOVA. Os Apêndices 10.3 e 10.4 apresentam os procedimentos que podem ser usados. Quanto ao exemplo da NCP, o valor p exato correspondente à estatística de teste $F = 9$ é 0,003.

À semelhança do que ocorre com outros procedimentos de teste de hipóteses, o critério do valor crítico também pode ser usado. Com $\alpha = 0{,}05$, o valor F crítico ocorre com uma área de 0,05 na cauda superior de uma distribuição F com 2 e 15 graus de liberdade. Na tabela de distribuição F, encontramos $F_{0{,}05} = 3{,}68$. Portanto, a regra de rejeição apropriada da cauda superior para o exemplo da NCP é:

$$\text{Rejeitar } H_0 \text{ se } F \geq 3{,}68$$

Com $F = 9$, rejeitamos H_0 e concluímos que as médias das três populações não são iguais. Um resumo do procedimento global para testar a igualdade de k médias populacionais é apresentado a seguir:

TESTE DA IGUALDADE DE k MÉDIAS POPULACIONAIS

H_0: $\mu_1 = \mu_2 = \ldots = \mu_k$

H_a: Nem todas as médias populacionais são iguais

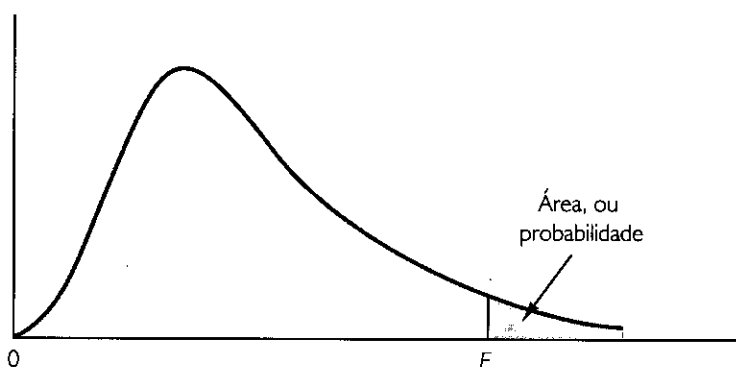
ESTATÍSTICA DE TESTE

$$F = \frac{\text{MSTR}}{\text{MSE}}$$

REGRA DE REJEIÇÃO

Critério do valor p:	Rejeitar H_0 se o valor $p \leq \alpha$
Critério do valor crítico:	Rejeitar H_0 se $F \geq F_\alpha$

em que o valor de F_α baseia-se em uma distribuição F com $k - 1$ graus de liberdade no numerador e $n_T - k$ graus de liberdade no denominador.

Tabela 10.4 Valores selecionados da tabela de distribuição F

Graus de Liberdade do Denominador	Área da Cauda Superior	Graus de Liberdade do Numerador 1	2	3	4	5
10	0,10	3,29	2,92	2,73	2,61	2,52
	0,05	4,96	4,10	3,71	3,48	3,33
	0,025	6,94	5,46	4,83	4,47	4,24
	0,01	10,04	7,56	6,55	5,99	5,64
15	0,10	3,07	2,70	2,49	2,36	2,27
	0,05	4,54	3,68	3,29	3,06	2,90
	0,025	6,20	4,77	4,15	3,80	3,58
	0,01	8,68	6,36	5,42	4,89	4,56
20	0,10	2,97	2,59	2,38	2,25	2,16
	0,05	4,35	3,49	3,10	2,87	2,71
	0,025	5,87	4,46	3,86	3,51	3,29
	0,01	8,10	5,85	4,94	4,43	4,10
25	0,10	2,92	2,53	2,32	2,18	2,09
	0,05	4,24	3,39	2,99	2,76	2,60
	0,025	5,69	4,29	3,69	3,35	3,13
	0,01	7,77	5,57	4,68	4,18	3,85
30	0,10	2,88	2,49	2,28	2,14	2,05
	0,05	4,17	3,32	2,92	2,69	2,53
	0,025	5,57	4,18	3,59	3,25	3,03
	0,01	7,56	5,39	4,51	4,02	3,70

Nota: A Tabela 4 do Apêndice B é uma tabela mais completa.

A Tabela ANOVA

Os resultados dos cálculos anteriores podem ser exibidos convenientemente em uma tabela denominada tabela de análise de variância, ou **tabela ANOVA**.[4] A Tabela 10.5 é a tabela de análise de variância correspondente ao exemplo da National Computer Products. A soma de quadrados associada à fonte de variação que recebe o rótulo de "Total" denomina-se *soma total dos quadrados* – SST (*total sum of squares*). Observe que os resultados correspondentes ao exemplo da NCP apresentam SST = SSTR + SSE, e que os graus de liberdade associados a essa soma total de quadrados é a soma dos graus de liberdade associados com a estimativa de σ^2 entre tratamentos e com a estimativa de σ^2 dentro de tratamentos.

Tabela 10.5 Tabela de análise de variância do exemplo da NCP

Fonte de Variação	Soma dos Quadrados	Graus de Liberdade	Quadrado Médio	F
Tratamentos	516	2	258,00	9,00
Erro	430	15	28,67	
Total	946	17		

Pode-se imaginar a análise de variância como um procedimento estatístico para dividir a soma total dos quadrados em componentes distintos.

Destacamos que a soma total dos quadrados (SST) dividida por seus graus de liberdade $n_T - 1$ é a variância amostral global que seria obtida se tratássemos o conjunto inteiro de 18 observações como um conjunto de dados. Quando se tem o conjunto de dados inteiro como uma única amostra, a fórmula para calcular a soma total dos quadrados, SST, é:

$$\text{SST} = \sum_{j=1}^{k}\sum_{i=1}^{n_j}(x_{ij} - \bar{\bar{x}})^2 \tag{10.22}$$

Pode-se demonstrar que os resultados que observamos na tabela de análise de variância correspondentes ao exemplo da NCP também se aplicam a outros problemas. Ou seja,

$$\text{SST} = \text{SSTR} + \text{SSE} \tag{10.23}$$

Em outras palavras, a SST pode ser dividida em duas somas de quadrados: a soma de quadrados dos tratamentos e a soma de quadrados dos erros. Note também que os graus de liberdade correspondentes a SST, $n_T - 1$, podem ser divididos nos graus de liberdade correspondentes a SSTR, $k - 1$, e nos graus de liberdade correspondentes a SSE, $n_T - k$. A análise de variância pode ser vista como um processo de **partição** da soma total dos quadrados e os graus de liberdade em suas fontes correspondentes: tratamentos e erro. Dividir a soma dos quadrados pelos graus de liberdade apropriados produzirá as estimativas de variância e o valor F que são usados para testar a hipótese de médias populacionais iguais.

Resultados de Computador para a Análise de Variância

Em virtude da ampla disponibilidade de pacotes de software estatístico, os cálculos da análise de variância com tamanhos de amostra grandes ou com um número grande de populações podem ser executados facilmente. Na Figura 10.6, apresentamos a saída de dados (*output*) correspondente ao exemplo da NCP obtida pelo software Minitab. A primeira parte da saída de dados do software contém o familiar formato da tabela ANOVA. Comparando a Figura 10.6 com a Tabela 10.5, vemos que a mesma informação está disponível, não obstante alguns cabeçalhos serem ligeiramente diferentes. O cabeçalho Source (Fonte) é usado para a coluna Source of Variation (Fonte de Variação), e Factor (Fator) identifica a linha Treatments (Tratamentos). As colunas Sum of Squares (Soma de Quadrados) e Degrees of Freedom (Graus de Liberdade) estão permutadas, e o valor p é fornecido para o teste F. Dessa forma, com o nível de significância $\alpha = 0{,}05$, rejeitamos H_0 porque o valor $p = 0{,}003 \leq \alpha = 0{,}05$.

Observe que depois da tabela de análise de variância (ANOVA), a saída de dados contém os respectivos tamanhos de amostra, as médias amostrais e os desvios padrão. Além disso, o Minitab produz uma

[4] NT: ANOVA – Sigla de *analysis of variance* (análise de variância).

imagem que exibe as estimações por intervalo individuais com 95% de confiança de cada média populacional. Para desenvolver essas estimações por intervalo de confiança, o Minitab usa o MSE como estimativa de σ^2. Desse modo, a raiz quadrada de MSE produz a melhor estimativa do desvio padrão σ da população. Essa estimativa de σ na saída de dados de computador é Pooled StDev (Desvio Padrão Agrupado); ele é igual a 5,354. Para apresentarmos uma ilustração de como essas estimações por intervalo são desenvolvidas, calcularemos uma estimação por intervalo com 95% de confiança da média populacional correspondente à fábrica de Atlanta.

Figura 10.6 Saída de dados do Minitab para a análise de variância da NCP

```
Analysis of Variance
Source     DF        SS       MS        F        P
Factor      2     516.0    258.0     9.00    0.003
Error      15     430.0     28.7
Total      17     946.0
                                  Individual 95% CIs For Mean
                                  Based on Pooled StDev
 Level      N      Mean    StDev  ---+---------+---------+---------+---
Atlanta     6    79.000    5.831                      (-----*-----)
Dallas      6    74.000    4.472            (-----*-----)
Seattle     6    66.000    5.657   (-----*-----)
                                  ---+---------+---------+---------+---
Pooled StDev =    5.354           63.0      70.0      77.0      84.0
```

Do estudo de estimação por intervalo que realizamos no Capítulo 8, sabemos que a forma geral de uma estimação por intervalo de uma média populacional é:

$$\bar{x} \pm t_{\alpha/2} \frac{s}{\sqrt{n}} \qquad \textbf{(10.24)}$$

em que σ é a estimativa do desvio padrão s da população. Na análise de variância, a melhor estimativa de σ é fornecida pela raiz quadrada do MSE ou do Desvio Padrão Agrupado (Pooled StDev); portanto, usamos o valor 5,354 para s na Equação 10.24. O grau de liberdade para $t_{\alpha/2}$ é 15, ou seja, o grau de liberdade associado à estimativa de σ^2 dentro de tratamentos. Portanto, com 95% de confiança, temos $t_{0,025}$ = 2,131 e

$$79 \pm 2{,}131 \frac{5{,}354}{\sqrt{6}} = 79 \pm 4{,}66$$

Assim, o intervalo de confiança individual de 95% para a fábrica de Atlanta abrange 79 – 4,66 = 74,34 a 79 + 4,66 = 83,66. Uma vez que os tamanhos de amostra são iguais para o exemplo da NCP, os intervalos de confiança individuais para as fábricas de Dallas e Seattle também são construídos adicionando-se e subtraindo-se 4,66 de cada média amostral. Desse modo, na imagem produzida pelo Minitab, notamos que as larguras dos intervalos de confiança são idênticas.

Exercícios

Métodos

AUTOTESTE

27. Cinco observações foram selecionadas de três populações. Os dados obtidos são os seguintes:

Observação	Amostra 1	Amostra 2	Amostra 3
1	32	44	33
2	30	43	36
3	30	44	35
4	26	46	36
5	32	48	40
Média da amostra	30	45	36
Variância da amostra	6,00	4,00	6,50

a. Calcule a estimativa de σ^2 entre tratamentos
b. Calcule a estimativa de σ^2 dentro de tratamentos.
c. Com o nível de significância $\alpha = 0{,}05$, podemos rejeitar a hipótese nula de que as médias das três populações são iguais?
d. Crie a tabela ANOVA para esse problema.

28. Quatro observações foram selecionadas de cada uma de três diferentes populações. Os dados obtidos são os seguintes:

Observação	Amostra 1	Amostra 2	Amostra 3
1	165	174	169
2	149	164	154
3	156	180	161
4	142	158	148
Média da amostra	153	169	158
Variância da amostra	96,67	97,33	82,00

a. Calcule a estimativa de σ^2 entre tratamentos
b. Calcule a estimativa de σ^2 dentro de tratamentos.
c. Com o nível de significância $\alpha = 0{,}05$, podemos rejeitar a hipótese nula de que as médias das três populações são iguais? Explique.
d. Crie a tabela ANOVA para esse problema.

29. Amostras foram selecionadas de três populações. Os dados obtidos são os seguintes:

	Amostra 1	Amostra 2	Amostra 3
	93	77	88
	98	87	75
	107	84	73
	102	95	84
		85	75
		82	
$\bar{x}_j$	100	85	79
s_j^2	35,33	35,60	43,50

a. Calcule a estimativa de σ^2 entre tratamentos.
b. Calcule a estimativa de σ^2 dentro de tratamentos.
c. Com o nível de significância $\alpha = 0{,}05$, podemos rejeitar a hipótese nula de que as médias das três populações são iguais? Explique.
d. Crie a tabela ANOVA para esse problema.

30. Uma amostra aleatória de 16 observações foi selecionada de cada uma de quatro diferentes populações. Uma parte da tabela ANOVA é apresentada a seguir:

Fonte de Variação	Soma dos Quadrados	Graus de Liberdade	Quadrado Médio	F
Tratamentos			400	
Erro				
Total	1.500			

a. Preencha os lançamentos que faltam na tabela ANOVA.
b. Com o nível de significância $\alpha = 0{,}05$, podemos rejeitar a hipótese nula de que as médias das quatro populações são iguais?

31. Amostras aleatórias de 25 observações foram selecionadas de cada uma de três diferentes populações. Para esses dados, a SSTR = 120 e a SSE = 216.

 a. Crie a tabela ANOVA para esse problema.
 b. Com o nível de significância $\alpha = 0,05$, podemos rejeitar a hipótese nula de que as médias das três populações são iguais?

Aplicações

AUTOTESTE

32. A fim de testar se o tempo médio necessário para misturar um lote de materiais é o mesmo para máquinas produzidas por três diferentes fabricantes, a Jacobs Chemical Company obteve os seguintes dados sobre o tempo (em minutos) necessário para misturar os materiais. Use esses dados para testar se o tempo médio populacional para misturar um lote de materiais difere em relação aos três fabricantes. Use $\alpha = 0,05$.

Fabricante		
1	2	3
20	28	20
26	26	19
24	31	23
22	27	22

33. O Texas Transportation Institute, da Texas A&M University, realizou uma pesquisa para determinar o número de horas por ano que os motoristas gastavam no trânsito. Das 75 áreas urbanas estudadas, a mais congestionada foi Los Angeles, onde os motoristas gastavam em média 90 horas por ano (*U.S. News & World Report*, 13 de outubro de 2003). Entre outras áreas urbanas congestionadas contavam-se Denver, Miami e São Francisco. Suponha que dados amostrais de seis motoristas de cada uma dessas cidades apresentem o seguinte número de horas gastas por ano no trânsito:

Denver	Miami	São Francisco
70	66	65
62	70	62
71	55	74
58	65	69
57	56	63
66	66	75

 a. Calcule a média amostral de horas gastas por ano correspondente a cada uma dessas áreas urbanas.
 b. Usando $\alpha = 0,05$, teste as diferenças de significância entre a média populacional de tempo gasto correspondente a cada uma dessas três áreas urbanas. Qual é o valor p? Qual é a sua conclusão?

34. Nova York, Boston e o Vale do Silício na Califórnia estão entre as regiões que apresentam os maiores salários no setor de tecnologia nos Estados Unidos (*USA Today*, 28 de fevereiro de 2002). Os dados amostrais seguintes apresentam os salários anuais individuais expressos em milhares de dólares.

ARQUIVO DA INTERNET

Technology

Nova York	Boston	Vale do Silício
82	85	82
79	80	91
72	74	94
89	78	88
79	75	85
85	80	
	86	
	74	

Use $\alpha = 0,05$ e teste a diferença de significância entre a média populacional de salários do setor de tecnologia correspondentes a essas três localidades. Qual é o valor p? Qual é a sua conclusão? Se existe uma diferença, qual localidade parece ter a média de salário mais elevada para o setor de tecnologia?

35. Um estudo divulgado no *Journal of Small Business Management* concluiu que as pessoas que trabalham como autônomos enfrentam maior grau de estresse no trabalho do que as pessoas que não são autônomas. Nesse estudo, o estresse no trabalho foi avaliado de acordo com uma escala de 15 itens idealizados para medir vários aspectos referentes a ambigüidade e conflito de cargos. As avaliações referentes a cada um dos 15 itens foram feitas com base em uma escala de 1 a 5 que indicava opções de resposta

que variavam de forte concordância a forte discordância. A soma das avaliações referentes aos 15 itens correspondentes a cada indivíduo pesquisado situa-se entre 15 e 75, e os valores mais elevados indicam maior grau de estresse no trabalho. Suponha que uma abordagem idêntica, usando uma escala de 20 itens com opções de resposta de 1 a 5, seja usada para avaliar o grau de estresse no trabalho de 15 agentes imobiliários, 15 arquitetos e 15 corretores da bolsa selecionados aleatoriamente.

Agente Imobiliário	Arquiteto	Corretor da Bolsa
81	43	65
48	63	48
68	60	57
69	52	91
54	54	70
62	77	67
76	68	83
56	57	75
61	61	53
65	80	71
64	50	54
69	37	72
83	73	65
85	84	58
75	58	58

ARQUIVO DA INTERNET
Stress

Use $\alpha = 0{,}05$ para testar se há quaisquer diferenças significativas no grau de estresse no trabalho entre as três profissões.

36. A *Condé Nast Traveler* realiza uma pesquisa anual na qual os leitores classificam seus navios de cruzeiro preferidos. As avaliações fornecidas referem-se a navios pequenos (que transportam até 500 passageiros), navios de porte médio (que transportam de 500 a 1.500 passageiros) e navios grandes (que transportam, no mínimo, 1.500 passageiros). Os dados a seguir mostram as avaliações de serviço relativas a oito navios pequenos selecionados aleatoriamente, oito navios de porte médio selecionados aleatoriamente e oito navios grandes selecionados aleatoriamente. Todos os navios são avaliados em uma escala de 100 pontos, e os valores mais elevados indicam melhor serviço (*Condé Nast Traveler*, fevereiro de 2003).

Navios Pequenos		Navios de Porte Médio		Navios Grandes	
Nome	Avaliação	Nome	Avaliação	Nome	Avaliação
Hanseactic	90,5	Amsterdam	91,1	Century	89,2
Mississippi Queen	78,2	Crystal Symphony	98,9	Disney Wonder	90,2
Philae	92,3	Maasdam	94,2	Enchantment of the Seas	85,9
Royal Clipper	95,7	Noordam	84,3	Grand Princess	84,2
Seabourn Pride	94,1	Royal Princess	84,8	Infinity	90,2
Seabourn Spirit	100	Ryndam	89,2	Legend of the Seas	80,6
Silver Cloud	91,8	Statendam	86,4	Paradise	75,8
Silver Wind	95	Veendam	88,3	Sun Princess	82,3

ARQUIVO DA INTERNET
Ships

Use $\alpha = 0{,}05$ para testar se há quaisquer diferenças significativas na média de avaliações do serviço entre os três tamanhos de navios de cruzeiro.

Resumo

Neste capítulo, apresentamos procedimentos para comparar duas ou mais médias populacionais. Primeiramente, mostramos como fazer inferências sobre a diferença entre duas médias populacionais quando amostras aleatórias simples independentes são selecionadas. Consideramos o caso em que se pode supor que os desvios padrão populacionais σ_1 e σ_2 são conhecidos. A distribuição normal padrão z foi usada para desenvolver a estimação por intervalo e serviu como estatística de teste os para testes de hipóteses. Depois, consideramos o caso em que os desvios padrão populacionais eram desconhecidos e estimados pelos desvios padrão amostrais s_1 e s_2. Nesse caso, a distribuição t foi usada para desenvolver a estimação por intervalo e serviu como estatística de teste para os testes de hipóteses.

Discutimos, então, as inferências sobre a diferença entre duas médias populacionais para o projeto de amostras relacionadas (ou dependentes). No projeto de amostras relacionadas, cada elemento fornece um par de valores de dados, sendo um de cada população. A diferença entre os valores de dados emparelha-

dos é então usada na análise estatística. O projeto de amostras relacionadas geralmente é preferível ao projeto de amostras independentes porque o projeto de amostras relacionadas freqüentemente melhora a precisão da estimativa.

Finalmente, mostramos como a análise de variância pode ser usada para testar as diferenças entre três ou mais médias populacionais. O procedimento de análise de variância usa duas estimativas da variância populacional, σ^2. A razão entre essas duas estimativas (a estatística F) pode ser usada para produzir o valor p e para determinar se a hipótese nula de que as médias populacionais são iguais deve ou não ser rejeitada.

Glossário

Amostras aleatórias simples independentes Amostras selecionadas de duas (ou mais) populações, de tal forma que os elementos que compõem uma amostra são escolhidos independentemente dos elementos que compõem a outra amostra.

Amostras relacionadas (ou **dependentes**) Amostras nas quais cada valor de dados de uma amostra se relaciona com um valor de dados correspondente da outra amostra.

Análise de variância (ANOVA) Uma técnica estatística que pode ser usada para testar a hipótese de que três ou mais médias populacionais são iguais.

Distribuição *F* Uma distribuição que se baseia na razão de duas estimativas independentes da variância de uma população normal. A distribuição é usada em testes de hipóteses sobre a igualdade de k médias populacionais.

Tabela ANOVA Uma tabela utilizada para resumir os cálculos e os resultados da análise de variância. Ela contém colunas que exibem a fonte de variação, a soma dos quadrados, os graus de liberdade, o quadrado médio e o valor F.

Partição O processo de alocar a soma total de quadrados e graus de liberdade a vários componentes.

Fórmulas-Chave

Estimador por Ponto da Diferença entre Duas Médias Populacionais

$$\bar{x}_1 - \bar{x}_2 \tag{10.1}$$

Erro padrão de $\bar{x}_1$ e $\bar{x}_2$

$$\sigma_{\bar{x}_1-\bar{x}_2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \tag{10.2}$$

Estimação por Intervalo da Diferença entre as Médias de Duas Populações: σ_1 e σ_2 Conhecidos

$$\bar{x}_1 - \bar{x}_2 \pm z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \tag{10.4}$$

Estatística de Teste para Testes de Hipóteses sobre $\mu_1 - \mu_2$ Quando σ_1 e σ_2 São Conhecidos

$$z = \frac{(\bar{x}_1 - \bar{x}_2) - D_0}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \tag{10.5}$$

Estimação por Intervalo da Diferença entre Duas Médias Populacionais Quando σ_1 e σ_2 São Desconhecidos

$$\bar{x}_1 - \bar{x}_2 \pm t_{\alpha/2}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \tag{10.6}$$

Graus de Liberdade da Distribuição *t* Quando se Usam Duas Amostras Aleatórias Independentes

$$df = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{1}{n_1 - 1}\left(\frac{s_1^2}{n_1}\right)^2 + \frac{1}{n_2 - 1}\left(\frac{s_2^2}{n_2}\right)^2} \tag{10.7}$$

Estatística de Teste para Testes de Hipótese sobre $\mu_1 - \mu_2$ Quando σ_1 e σ_2 São Desconhecidos

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - D_0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \tag{10.8}$$

Estatística de Teste para Testes de Hipótese que Envolvem Amostras Relacionadas (ou Dependentes)

$$t = \frac{\bar{d} - \mu_d}{s_d/\sqrt{n}} \tag{10.9}$$

Média Amostral para o Tratamento j

$$\bar{x}_j = \frac{\sum_{i=1}^{n_j} x_{ij}}{n_j} \tag{10.10}$$

Variância Amostral para o Tratamento j

$$s_j^2 = \frac{\sum_{i=1}^{n_j} (x_{ij} - \bar{x}_j)^2}{n_j - 1} \tag{10.11}$$

Média Global da Amostra

$$\bar{\bar{x}} = \frac{\sum_{j=1}^{k} \sum_{i=1}^{n_j} x_{ij}}{n_T} \tag{10.12}$$

$$n_T = n_1 + n_2 + \cdots + n_k \tag{10.13}$$

Quadrado Médio dos Tratamentos

$$\text{MSTR} = \frac{\text{SSTR}}{k - 1} \tag{10.16}$$

Soma de Quadrados dos Tratamentos

$$\text{SSTR} = \sum_{j=1}^{k} n_j (\bar{x}_j - \bar{\bar{x}})^2 \tag{10.17}$$

Quadrado Médio dos Erros

$$\text{MSE} = \frac{\text{SSE}}{n_T - k} \tag{10.19}$$

Soma de Quadrados dos Erros

$$\text{SSE} = \sum_{j=1}^{k} (n_j - 1) s_j^2 \tag{10.20}$$

Estatística de Teste da Igualdade de k Médias Populacionais

$$F = \frac{\text{MSTR}}{\text{MSE}} \tag{10.21}$$

Soma Total dos Quadrados

$$\text{SST} = \sum_{j=1}^{k} \sum_{i=1}^{n_j} (x_{ij} - \bar{\bar{x}})^2 \tag{10.22}$$

Partição da Soma de Quadrados

$$\text{SST} = \text{SSTR} + \text{SSE} \tag{10.23}$$

Exercícios Suplementares

37. A Safegate Foods, Inc., está redesenhando a posição dos caixas em seus supermercados em todo o país e está considerando dois desenhos (*designs*). Testes para verificar o tempo de que os clientes necessitam para serem atendidos nos caixas foram realizados em duas lojas onde os dois novos sistemas foram instalados e os resultados são apresentados no seguinte resumo:

Sistema A	Sistema B
$n_1 = 120$	$n_2 = 100$
$\bar{x}_1 = 4{,}1$ minutos	$\bar{x}_2 = 3{,}4$ minutos
$\sigma_1 = 2{,}2$ minutos	$\sigma_2 = 1{,}5$ minuto

Teste, com o grau de significância 0,05, e determine se a média populacional dos tempos de atendimento nos caixas dos dois sistemas diferem. Qual sistema é preferível?

38. Os salários anuais iniciais de pessoas que têm os graus de bacharel e de mestrado em Ciências da Administração foram coletados em duas amostras aleatórias independentes. Use os dados apresentados a seguir para desenvolver uma estimação por intervalo de confiança de 90% do aumento dos salários iniciais que se pode esperar após a conclusão de um programa de mestrado.

Mestrado	Bacharelado
$n_1 = 60$	$n_2 = 80$
$\bar{x}_1$ = US\$ 45.000	$\bar{x}_2$ = US\$ 35.000
σ_1 = US\$ 4.000	σ_2 = US\$ 3.500

39. As câmeras fotográficas digitais de três megapixels são, tipicamente, as mais leves, mais compactas e mais fáceis de usar. Entretanto, se você planeja ampliar ou recortar as imagens, provavelmente estará disposto a gastar mais para obter um modelo com resolução maior. Os dados a seguir apresentam os preços amostrais de câmeras digitais de cinco megapixels e de três megapixels (*Consumer Reports Buying Guide*, 2004).

ARQUIVO DA INTERNET

Digital

Cinco megapixels		Três megapixels	
Modelo	Preço	Modelo	Preço
Nikon 5700	890	Kodak DX4330	280
Olympus C-5050	620	Canon A70	290
Sony DCS-F717	730	Sony DSC P8	370
Olympus C-5050	480	Minolta XI	400
Minolta 7Hi	1060	Sony DSC P72	310
HP 935	450	Nikon 3100	340
Pentax 550	540	Panasonic DMC-LC33	270
Canon S50	500	Pentax S	380
Kyocera TVS	890		
Minolta F300	440		

a. Forneça uma estimação por ponto das diferenças entre a média populacional de preços correspondentes aos dois tipos de câmera digital. Quais observações você é capaz de fazer sobre o preço do modelo de cinco megapixels de maior qualidade?

b. Desenvolva uma estimação por intervalo de 95% de confiança da diferença entre as duas médias populacionais de preços.

40. Os fundos mútuos são classificados como *load* ou *no-load*. Os *load funds* exigem que o investidor pague uma taxa inicial baseada em uma porcentagem da quantia investida no fundo mútuo. Os *no-load funds* não exigem essa taxa inicial. Alguns consultores financeiros argumentam que os *load mutual funds* podem valer a taxa extra, porque esses fundos rendem uma taxa média de retorno mais alta que os *no-load mutual funds*. Uma amostra de 30 *load mutual funds* e uma amostra de 30 *no-load mutual funds* foram selecionadas. Dados foram coletados sobre o rendimento anual dos fundos ao longo de um período de cinco anos. Os dados estão contidos no conjunto de dados (*data set*) intitulado Mutual. Os dados correspondentes aos cinco primeiros *load funds* e aos cinco primeiros *no-load funds* são os seguintes:

Fundos Mutual – *Load*	Rendimento	Fundos Mutual – No-*Load*	Rendimento
American National Growth	15,51	Amana Income Fund	13,24
Arch Small Cap Equity	14,57	Berger One Hundred	12,13
Bartlett Cap Basic	17,73	Columbia International Stock	12,17
Calvert World International	10,31	Dodge & Cox Balanced	16,06
Colonial Fund A	16,23	Evergreen Fund	17,61

ARQUIVO DA INTERNET

Mutual

a. Formule H_0 e H_a de forma que a rejeição de H_0 leve à conclusão de que os fundos mútuos *load* têm um retorno médio anual mais elevado ao longo de um período de cinco anos.

b. Use os fundos mútuos de 60 dias do conjunto de dados intitulado Mutual para realizar o teste de hipóteses. Qual é o valor p? Com $\alpha = 0{,}05$, qual é a sua conclusão?

41. A National Association of Home Builders publicou dados sobre o custo dos projetos mais populares de reforma de residências. Dados amostrais sobre o custo, em milhares de dólares para dois tipos de projetos de reforma, são os seguintes:

Cozinha	Quarto de casal	Cozinha	Quarto de casal
25,2	18,0	23,0	17,8
17,4	22,9	19,7	24,6
22,8	26,4	16,9	21,0
21,9	24,8	21,8	
19,7	26,9	23,6	

a. Desenvolva uma estimação por ponto da diferença entre a média populacional dos custos de reforma para os dois tipos de projeto.

b. Desenvolva um intervalo de confiança de 90% da diferença entre as duas médias populacionais.

42. Os preços típicos das casas de moradia para uma família no estado da Flórida são apresentados a seguir e correspondem a uma amostra de 15 regiões metropolitanas (*Naples Daily News*, 23 de fevereiro de 2003). Os dados estão expressos em milhares de dólares:

Região Metropolitana	Janeiro 2003	Janeiro 2002
Daytona Beach	117	96
Fort Lauderdale	207	169
Fort Myers	143	129
Fort Walton Beach	139	134
Gainesville	131	119
Jacksonville	128	119
Lakeland	91	85
Miami	193	165
Naples	263	233
Ocala	86	90
Orlando	134	121
Pensacola	111	105
Sarasota-Bradenton	168	141
Tallahassee	140	130
Tampa-St. Petersburg	139	129

ARQUIVO DA INTERNET

Florida

a. Use uma análise de amostras relacionadas (ou dependentes) para desenvolver uma estimação por ponto da média populacional do aumento de preços ao longo de um ano para as casas de moradia para uma família no estado da Flórida.

b. Desenvolva uma estimação por intervalo de confiança de 90% para a média populacional do aumento de preços ao longo de um ano para as casas de moradia para uma família no estado da Flórida.

c. Qual foi o aumento percentual no período de um ano?

43. A revista *Money* divulga os rendimentos percentuais e os índices de despesas correspondentes aos fundos de títulos e ações. Os dados a seguir são os índices de despesas de dez fundos de títulos mobiliários de média capitalização (*midcap*), de dez fundos de títulos mobiliários de pequena capitalização, de dez fundos de ações híbridos e de dez fundos de ações especiais (*Money*, março de 2003).

ARQUIVO DA INTERNET
Funds

Média Capitalização	Pequena Capitalização	Híbridos	Especiais
1,2	2,0	2,0	1,6
1,1	1,2	2,7	2,7
1,0	1,7	1,8	2,6
1,2	1,8	1,5	2,5
1,3	1,5	2,5	1,9
1,8	2,3	1,0	1,5
1,4	1,9	0,9	1,6
1,4	1,3	1,9	2,7
1,0	1,2	1,4	2,2
1,4	1,3	0,3	0,7

Use $\alpha = 0,05$ para testar se há alguma diferença significativa no índice médio de despesas entre os quatro tipos de fundos de ações.

44. Os compradores de veículos utilitários esportivos e picapes têm ampla variedade de escolha no mercado atual. Um dos fatores importantes para muitos compradores é o valor de revenda do veículo. A tabela a seguir apresenta o valor de revenda (%) de dez utilitários esportivos, dez picapes pequenas e dez picapes grandes depois de dois anos de uso (*Kiplinger's New Cars & Trucks 2000 Buyer's Guide*).

ARQUIVO DA INTERNET
Trucks

Utilitário Esportivo	Valor de Revenda	Picape Pequena	Valor de Revenda
Chevrolet Blazer LS	55	Chevrolet S-10 Extended Cab	46
Ford Explorer Sport	57	Dodge Dakota Club Cab Sport	53
GMC Yukon XL 1500	67	Ford Ranger XLT Regular Cab	48
Honda CR-V	65	Ford Ranger XLT Supercab	55
Isuzu VehiCross	62	GMC Sonoma Regular Cab	44
Jeep Cherokee Limited	57	Isuzu Hombre Spacecab	41
Mercury Mountaineer	59	Mazda B4000 SE Cab Plus	51
Nissan Pathfinder XE	54	Nissan Frontier XE Regular Cab	51
Toyota 4Runner	55	Toyota Tacoma Xtracab	49
Toyota RAV4	55	Toyota Tacoma Xtracab V6	50

Picape Grande	Valor de Revenda
Chevrolet K2500	60
Chevrolet Silverado 2500 Ext	64
Dodge Ram 1500	54
Dodge Ram Quad Cab 2500	63
Dodge Ram Regular Cab 2500	59
Ford F150 XL	58
Ford F350 Super Duty Crew Cab XL	64
GMC New Sierra 1500 Ext Cab	68
Toyota Tundra Access Cab Limited	53
Toyota Tundra Regular Cab	58

Com o nível de confiança $\alpha = 0,05$, teste se há alguma diferença significativa no valor médio de revenda correspondente aos três tipos de veículo.

45. A empresa Crowne Plaza Hotel and Resorts ofereceu preços especiais de fim de semana nos hotéis e estâncias de sua propriedade em todo o país. Uma amostra de 30 propriedades de três regiões do país forneceu os seguintes preços de quartos (*USA Today*, 14 de abril de 2000).

ARQUIVO DA INTERNET
Resorts

Oeste	Preço ($)	Sul	Preço ($)	Nordeste	Preço ($)
Albuquerque	89	Atlanta	105	Albany	89
Irvine	79	Dallas	80	Boston	139
Las Vegas	119	Greenville	79	Hartford	85
Los Angeles	99	Houston	79	New York	159
Palo Alto	109	Jackson	69	Philadelphia	99
Phoenix	149	Macon	69	Pittsfield	99
Portland	79	Miami	89	Providence	149
San Francisco	139	Orlando	119	Washington	159
San Jose	99	Richmond	109	White Plains	109
Seattle	119	Tampa	119	Worchester	124

Com o nível de significância $\alpha = 0,05$, teste se os preços médios são os mesmos nas três regiões.

46. A National Footbal League avalia os candidatos a jogador de acordo com a posição, em uma escala que varia de 5 a 9. As avaliações são interpretadas da seguinte maneira: 8–9 deve começar a jogar no primeiro ano, 7,0–7,9 está apto a começar a jogar, 6,0–6,9 integrará o time na posição de reserva, e 5,0–5,9 pode fazer parte do clube e contribuir. A tabela a seguir apresenta as classificações referentes a três posições de 40 candidatos a jogador na NFL (*USA Today*, 14 de abril de 2000). A posição em que o jogador joga parece ter algum efeito significativo sobre a avaliação?

ARQUIVO DA INTERNET
NFL

Wide Receiver[5]		Guard[6]		Offensive Tackle	
Nome	Classificação	Nome	Classificação	Nome	Classificação
Peter Warrick	9,0	Cosey Coleman	7,4	Chris Samuels	8,5
Plaxico Burress	8,8	Travis Claridge	7,0	Stockar McDougle	8,0
Sylvester Morris	8,3	Kaulana Noa	6,8	Chris McIngosh	7,8
Travis Taylor	8,1	Leander Jordan	6,7	Adrian Klemm	7,6
Laveranues Coles	8,0	Chad Clifton	6,3	Todd Wade	7,3
Dez White	7,9	Manula Savea	6,1	Marvel Smith	7,1
Jerry Porter	7,4	Ryan Johanningmeir	6,0	Michael Thompson	6,8
Ron Dugans	7,1	Mark Tauscher	6,0	Bobby Williams	6,8
Todd Pinkston	7,0	Blaine Saipaia	6,0	Darnell Alford	6,4
Dennis Northcutt	7,0	Richard Mercier	5,8	Terrance Beadles	6,3
Anthony Lucas	6,9	Damion McIntosh	5,3	Tutan Reyes	6,1
Darrell Jackson	6,6	Jeno James	5,5	Greg Robinson-Ran	6,0
Danny Farmer	6,5	Al Jackson	5,5		
Sherrod Gideon	6,4				
Trevor Gaylor	6,2				

Estudo de Caso 1 – Par, Inc.

A Par, Inc. é uma grande fábrica de equipamentos de golfe. A administração acredita que a participação de mercado da Par poderia ser aumentada com a introdução de uma bola de golfe resistente a cortes e de maior durabilidade. Portanto, uma equipe de pesquisa da Par investiga a produção de um novo revestimento de bolas de golfe projetado para resistir a cortes e produzir uma bola mais durável. Os testes com o revestimento têm sido promissores.

Um dos pesquisadores manifestou preocupação acerca do efeito do novo revestimento sobre as distâncias de arremesso (*driving distances*). A Par gostaria que a bola com o novo revestimento atingisse distâncias de arremesso comparáveis às do modelo de bola de golfe atual. Para comparar as distâncias de arremesso das duas bolas, 40 bolas do modelo novo e do modelo antigo foram submetidas a testes de distância. Os testes foram executados com uma máquina de disparo mecânico a fim de que quaisquer diferenças entre as distâncias médias obtidas pelos dois modelos pudessem ser atribuídas a uma diferença nos dois modelos. Os resultados dos testes, sendo as distâncias medidas em jardas, de acordo com a menor distância percorrida, são apresentados a seguir. Esses dados estão disponíveis na página do livro na internet.

ARQUIVO DA INTERNET
Golf

Modelo		Modelo		Modelo		Modelo	
Atual	Novo	Atual	Novo	Atual	Novo	Atual	Novo
264	277	270	272	263	274	281	283
261	269	287	259	264	266	274	250
267	263	289	264	284	262	273	253
272	266	280	280	263	271	263	260
258	262	272	274	260	260	275	270
283	251	275	281	283	281	267	263
258	262	265	276	255	250	279	261
266	289	260	269	272	263	274	255
259	286	278	268	266	278	276	263
270	264	275	262	268	264	262	279

[5] NT: *Wide Receiver* – O jogador que recebe os lançamentos em linha avançada para conseguir o máximo de jardas à frente (*futebol norte-americano*).

[6] NT: *Guard* – A primeira linha de ataque é composta pelos jogadores mais corpulentos do time, sendo sua função bloquear a defesa adversária. Essa primeira linha é composta por um *center*, dois *guards*, dois *offensive tackles* e um *tight end*(*futebol norte-americano*).

Relatório Administrativo

1. Formule e apresente o fundamento lógico para o teste de hipóteses que a Par poderia usar para comparar as distâncias de arremesso das bolas de golfe atual e nova.
2. Analise os dados para fornecer a conclusão do teste de hipóteses. Qual é o valor p de seu teste? Qual é a sua recomendação à Par, Inc.?
3. Apresente sumários estatísticos dos dados correspondentes a cada modelo.
4. Qual é o intervalo de confiança de 95% da média populacional de cada modelo, e qual é o intervalo de confiança de 95% da diferença entre as médias das duas populações?
5. Você vê a necessidade de tamanhos de amostra maiores e de mais testes com as bolas de golfe? Discuta.

Estudo de Caso 2 – Wentworth Medical Center

Como parte de um estudo de longo prazo de pessoas com idades a partir de 65 anos, sociólogos e médicos do Wentworth Medical Center, localizado na região norte de Nova York, investigaram a relação entre localização geográfica e depressão. Uma amostra de 60 pessoas, todas com saúde relativamente boa, foi selecionada; 20 indivíduos residiam na Flórida, 20 residiam em Nova York e 20 residiam na Carolina do Norte. Cada um dos indivíduos integrantes da amostra foi submetido a um exame padronizado para medir a depressão. Os dados coletados são apresentados a seguir; pontuações mais altas no exame indicam maiores níveis de depressão. Esses dados estão disponíveis no arquivo intitulado Medical1 no site.

Uma segunda parte do estudo considerou a relação entre localização geográfica e depressão para pessoas com idades a partir de 65 anos que apresentavam problemas crônicos de saúde, por exemplo, artrite, hipertensão e/ou moléstia cardíaca. Uma amostra de 60 pessoas com essas condições foi identificada. Novamente, 20 residiam na Flórida, 20 residiam em Nova York e 20 residiam na Carolina do Norte. Os níveis de depressão registrados nesse estudo são apresentados a seguir. Os dados estão disponíveis no site.

ARQUIVO DA INTERNET Medical1

ARQUIVO DA INTERNET Medical2

Dados da Medical1			Dados da Medical2		
Flórida	Nova York	Carolina do Norte	Flórida	Nova York	Carolina do Norte
3	8	10	13	14	10
7	11	7	12	9	12
7	9	3	17	15	15
3	7	5	17	12	18
8	8	11	20	16	12
8	7	8	21	24	14
8	8	4	16	18	17
5	4	3	14	14	8
5	13	7	13	15	14
2	10	8	17	17	16
6	6	8	12	20	18
2	8	7	9	11	17
6	12	3	12	23	19
6	8	9	15	19	15
9	6	8	16	17	13
7	8	12	15	14	14
5	5	6	13	9	11
4	7	3	10	14	12
7	7	8	11	13	13
3	8	11	17	11	11

Relatório Administrativo

1. Use estatística descritiva para sintetizar os dados dos dois estudos. Quais são suas observações preliminares a respeito das pontuações obtidas no exame relativo à depressão?
2. Use a análise de variância em ambos os conjuntos de dados. Estabeleça as hipóteses a serem testadas em cada caso. Quais são suas conclusões?

3. Use inferências a respeito das médias de tratamento individuais, quando for apropriado. Quais são suas conclusões?
4. Discuta possíveis extensões desse estudo ou outras análises que você acha que possam ser úteis.

Estudo de Caso 3 – Remuneração de Profissionais de ID

Durante os últimos dez anos, a *Industrial Distribution* acompanhou a remuneração recebida por profissionais da distribuição industrial (ID). Os resultados obtidos dos 358 entrevistados na Annual Salary Survey (Pesquisa Anual de Salários) de 1997 mostraram que 27% das pessoas trabalham em empresas com níveis de venda superiores a US$ 40 milhões, e o profissional de ID típico trabalha para empresas de US$ 12 milhões. Aqueles que trabalham para empresas de pequeno a médio portes (entre US$ 6 e US$ 20 milhões) relatam salários mais elevados que aqueles que trabalham para empresas de maior porte. Os empregados que ganham menos trabalham para empresas com vendas inferiores a US$ 1 milhão. O vendedor externo típico ganhou US$ 50 mil em 1996 e o vendedor interno típico ganhou apenas US$ 30 mil (*Industrial Distribution*, novembro de 1997). Suponha que uma associação local de profissionais de ID da região da Grande São Francisco tenha realizado uma pesquisa entre seus membros para estudar a relação, se houver, entre os anos de experiência profissional e os salários das pessoas que ocupam funções de vendas externas e internas. Na pesquisa, os entrevistados foram solicitados a especificar um de três níveis de experiência profissional: baixa (1 a 10 anos), média (11 a 20 anos) e elevada (21 anos ou mais). Apresentamos a seguir uma parte dos dados obtidos. O conjunto de dados (*data set*) completo, o qual consiste em 120 observações, está disponível no arquivo intitulado IDSalary, no site.

ARQUIVO DA INTERNET
ID Salary

Observação	Salário $	Função	Experiência
1	28.938	Interna	Médio
2	27.694	Interna	Médio
3	45.515	Externa	Baixo
4	27.031	Interna	Médio
5	37.283	Externa	Baixo
6	32.718	Interna	Baixo
7	54.081	Externa	Elevado
8	23.621	Interna	Baixo
9	47.835	Externa	Elevado
10	29.768	Interna	Médio
•	•	•	•
•	•	•	•
•	•	•	•
115	33.080	Interna	Elevado
116	53.702	Externa	Médio
117	58.131	Externa	Médio
118	32.788	Interna	Elevado
119	28.070	Interna	Médio
120	35.259	Externa	Baixo

Relatório Administrativo

1. Use estatística descritiva para sintetizar os dados.
2. Desenvolva uma estimação por intervalo de confiança de 95% do salário anual médio de todos os vendedores, independentemente dos anos de experiência profissional.
3. Desenvolva uma estimação por intervalo de confiança de 95% da média salarial dos vendedores externos. Compare seus resultados com o valor nacional relatado pela *Industrial Distribution*.
4. Desenvolva uma estimação por intervalo de confiança de 95% da média salarial dos vendedores internos. Compare seus resultados com o valor nacional relatado pela *Industrial Distribution*.
5. Ignorando os anos de experiência, desenvolva uma estimação por intervalo de confiança de 95% da diferença média entre o salário anual dos vendedores externos e o salário anual médio dos vendedores internos. Qual é a sua conclusão?

6. Use a análise de variância para testar se há diferenças significativas em conseqüência da função exercida. Use o nível de significância 0,05 e, por enquanto, ignore o efeito dos anos de experiência.
7. Use a análise de variância para testar se há diferenças significativas em razão dos anos de experiência. Use o nível de significância 0,05 e, por enquanto, ignore o efeito da função exercida.
8. Com o nível de significância 0,05, teste se há diferenças significativas em decorrência da função exercida, dos anos de experiência e da interação. Use inferências sobre as médias de tratamento, quando apropriado.

Apêndice 10.1 – Inferências sobre Duas Populações com o Minitab

Descrevemos o uso do Minitab para desenvolver estimações por intervalo e realizar testes de hipóteses a respeito das diferenças entre duas médias populacionais e a diferença entre duas proporções populacionais. O Minitab fornece tanto os resultados de estimação por intervalo como de testes de hipóteses dentro do mesmo módulo. Desse modo, o procedimento do Minitab é o mesmo para ambos os tipos de inferências. Nos exemplos que apresentamos a seguir demonstraremos a estimação por intervalo e o teste de hipóteses para os mesmos dois exemplos. Observamos que o Minitab não apresenta uma rotina para a realização de inferências sobre duas médias populacionais quando os desvios padrão populacionais σ_1 e σ_2 são conhecidos.

Diferença entre Duas Médias Populacionais Quando σ_1 e σ_2 São Desconhecidos

ARQUIVO DA INTERNET

CheckAcct

Usaremos os dados do exemplo de saldos de conta corrente apresentados na Seção 10.2. Os saldos de conta corrente na filial de Cherry Grove estão na coluna C1 e os saldos de conta corrente da filial de Beechmont estão na coluna C2. Nesse exemplo, usaremos o procedimento 2-Sample *t* do Minitab para produzir uma estimação por intervalo de confiança de 95% da diferença entre as médias populacionais correspondentes aos saldos de contas correntes dos dois bancos filiais. A saída de dados (*output*) do procedimento também fornece o valor *p* do teste de hipóteses: H_0: $\mu_1 - \mu_2 = 0$ contra H_a: $\mu_1 - \mu_2 \neq 0$. As etapas a seguir são necessárias para se executar o procedimento:

Etapa 1. Selecione o menu **Stat**
Etapa 2. Escolha a opção **Basic Statistics**
Etapa 3. Escolha a opção **2-Sample t**
Etapa 4. Quando a caixa de diálogo 2-Sample t (Test and Confidence Interval) aparecer:
  Selecione **Samples in different columns**
  Digite C1 na caixa **First**
  Digite C2 na caixa **Second**
  Selecione **Options**
Etapa 5. Quando a caixa de diálogo 2-Sample t – Options aparecer:
  Digite 95 na caixa **Confidence level**
  Digite 0 na caixa **Test difference**
  Digite not qual (não igual) na caixa **Alternative**
  Dê um clique em **OK**
Etapa 6. Dê um clique em **OK**

A estimação por intervalo de confiança de 95% varia de US$ 37 a US$ 193, conforme descrevemos na Seção 10.2. O valor $p = 0{,}005$ mostra que a hipótese nula de médias populacionais iguais pode ser rejeitada ao nível de significância $\alpha = 0{,}01$. Em outras aplicações, a etapa 5 pode ser modificada para produzir diferentes níveis de confiança, diferentes valores hipotéticos e diferentes formas das hipóteses.

Diferença entre Duas Médias Populacionais com Amostras Relacionadas (ou Dependentes)

ARQUIVO DA INTERNET

Matched

Usamos os dados dos tempos de produção apresentados na Tabela 10.2 para ilustrar o procedimento de amostras relacionadas. Os tempos de conclusão correspondentes ao método 1 foram introduzidos na coluna C1 e os tempos de conclusão correspondentes ao método 2 foram introduzidos na coluna C2. As etapas do Minitab para as amostras relacionadas são as seguintes:

Etapa 1. Selecione o menu **Stat**
Etapa 2. Escolha a opção **Basic Statistics**
Etapa 3. Escolha a opção **Paired t**
Etapa 4. Quando a caixa de diálogo Paired t (Test and Confidence Interval) aparecer:
Selecione **Samples in columns**
Digite C1 na caixa **First sample**
Digite C2 na caixa **Second sample**
Selecione **Options**
Etapa 5. Quando a caixa de diálogo Paired t – Options aparecer:
Digite 95 na caixa **Confidence level**
Digite 0 na caixa **Test mean**
Digite not equal (não igual) na caixa **Alternative**
Dê um clique em **OK**
Etapa 6. Dê um clique em **OK**

A etapa 5 pode ser modificada para produzir diferentes níveis de confiança, diferentes valores hipotéticos e diferentes formas das hipóteses.

Apêndice 10.2 – Inferências sobre Duas Populações com o Excel

Descrevemos o uso do Excel para realizar testes de hipóteses a respeito da diferença entre duas médias populacionais.* Iniciamos com inferências a respeito da diferença entre a média de duas populações quando os desvios padrão populacionais, σ_1 e σ_2 são conhecidos.

ARQUIVO DA INTERNET
ExamScores

Diferença entre Duas Médias Populacionais Quando σ_1 e σ_2 São Conhecidos

Usaremos as notas (pontuações) de exame referentes aos dois centros de ensino discutidas na Seção 10.1. O rótulo Centro A está na célula A1 e o rótulo Centro B está na célula B1. As notas de exame correspondentes ao Centro A estão nas células A2:A31, e as notas de exame correspondentes ao Centro B estão nas células B2:B41. Presume-se que os desvios padrão populacionais sejam conhecidos, sendo $\sigma_1 = 10$ e $\sigma_2 = 10$. A rotina do Excel solicitará a entrada de variâncias, as quais são $\sigma_1^2 = 100$ e $\sigma_2^2 = 100$. As etapas seguintes podem ser usadas para realizar um teste de hipóteses sobre a diferença entre as duas médias populacionais.

Etapa 1. Selecione o menu **Ferramentas**
Etapa 2. Escolha a opção **Análise de Dados**
Etapa 3. Quando a caixa de diálogo Análise de Dados aparecer:
Escolha a opção **Teste-z: Duas Amostras para Médias**
Dê um clique em **OK**
Etapa 4. Quando a caixa de diálogo Teste-z: Duas Amostras para Médias aparecer:
Digite A1:A31 na caixa **Intervalo da variável 1**
Digite B1:B41 na caixa **Intervalo da variável 2**
Digite 0 na caixa **Hipótese de Diferença de Média**
Digite 100 na caixa **Variância da variável 1**
Digite 100 na caixa **Variância da variável 2**
Marque a opção **Rótulos**
Digite 0,05 na caixa **Alfa**
Selecione **Intervalo de Saída** e digite C1 na caixa
Dê um clique em **OK**

*As ferramentas de análise de dados do Excel oferecem procedimentos de teste de hipóteses para a diferença entre duas médias populacionais. Entretanto, não há nenhuma rotina no Excel para estimação por intervalo da diferença entre duas médias populacionais.

ARQUIVO DA INTERNET
Software Test

Diferença entre Duas Médias Populacionais Quando σ_1 e σ_2 São Desconhecidos

Usamos os dados do estudo sobre testes de software apresentados na Tabela 10.1. Os dados já foram inseridos em uma planilha do Excel, com o rótulo Atual na célula A1 e o rótulo Novo na célula B1. Os tempos de conclusão relativos ao uso da tecnologia atual estão nas células A2:A13 e os tempos de conclusão relativos ao uso do novo software estão nas células B2:B13. As etapas a seguir podem ser usadas para se realizar um teste de hipóteses a respeito da diferença entre duas médias populacionais quando σ_1 e σ_2 são desconhecidos.

Etapa 1. Selecione o menu **Ferramentas**
Etapa 2. Escolha a opção **Análise de Dados**
Etapa 3. Quando a caixa de diálogo Análise de Dados aparecer:
Escolha a opção **Teste-t: Duas Amostras Presumindo Variâncias Diferentes**
Dê um clique em **OK**
Etapa 4. Quando a caixa de diálogo Teste-t: Duas Amostras Presumindo Variâncias Diferentes aparecer:
Digite A1:A13 na caixa **Intervalo da variável 1**
Digite B1:B13 na caixa **Intervalo da variável 2**
Digite 0 na caixa **Hipótese de Diferença de Média**
Marque a opção **Rótulos**
Digite 0,05 na caixa **Alfa**
Selecione **Intervalo de Saída** e digite C1 na caixa
Dê um clique em **OK**

ARQUIVO DA INTERNET
Matched

Diferenças entre Duas Médias Populacionais com Amostras Relacionadas (ou Dependentes)

Usamos como ilustração os tempos de conclusão das amostras relacionadas da Tabela 10.2. Os dados foram introduzidos em uma planilha com o rótulo Método 1 na célula A1 e o rótulo Método 2 na célula B1. Os tempos de conclusão correspondentes ao método 1 estão nas células A2:A7 e os tempos de conclusão correspondentes ao método 2 estão nas células B2:B7. O procedimento do Excel utiliza as etapas descritas anteriormente com respeito ao Teste-*t*, excetuando-se que o usuário escolhe a ferramenta de análise de dados **Teste-t: Duas Amostras em Par para Médias** na etapa 3. O intervalo da variável 1 é A1:A7 e o intervalo da variável 2 é B1:B7.

ARQUIVO DA INTERNET
NCP

Apêndice 10.3 – Análise de Variância com o Minitab

Para ilustrar como o Minitab pode ser usado para testar a igualdade de *k* médias populacionais, mostramos como testar se a média das notas obtidas no exame é idêntica em cada fábrica, no exemplo da National Computer Products apresentado na Seção 10.4. Os dados da média das notas obtidas no exame foram inseridos nas três primeiras colunas de uma planilha do Minitab; a coluna 1 está rotulada como Atlanta, a coluna 2 está rotulada como Dallas e a coluna 3 está rotulada como Seattle.

As etapas a seguir produzem a saída de dados do Minitab da Figura 10.6.

Etapa 1. Selecione o menu **Stat**
Etapa 2. Escolha a opção **ANOVA**
Etapa 3. Escolha a opção **One-way (Unstacked)**
Etapa 4. Quando a caixa de diálogo One-Way Analysis of variance aparecer:
Digite C1-C3 na caixa **Responses (in separate columns)**
Dê um clique em **OK**

ARQUIVO DA INTERNET
NCP

Apêndice 10.4 – Análise de Variância com o Excel

Para ilustrar como o Excel pode ser usado para testar a igualdade de *k* médias populacionais correspondentes a ambos os casos, mostramos como testar se a média das notas obtidas no exame é idêntica em cada uma das fábricas, no exemplo da National Computer Products apresentado na Seção 10.4. Os dados sobre

as notas de exame foram inseridos nas linhas 2 a 7 das colunas B, C e D, como é exposto na Figura 10.7; note que as células da linha 1 têm os rótulos Atlanta, Dallas e Seattle. As etapas a seguir são usadas para se obter a saída de dados (*output*) mostrada nas células A9:G23; a parte ANOVA desta saída de dados corresponde à tabela ANOVA mostrada na Tabela 10.5.

Etapa 1. Selecione o menu **Ferramentas**
Etapa 2. Escolha a opção **Análise de Dados**

Figura 10.7 Solução do Excel para o exemplo de análise de variância da NCP

	A	B	C	D	E	F	G	H
1	**Observação**	**Atlanta**	**Dallas**	**Seattle**				
2	1	85	71	59				
3	2	75	75	64				
4	3	82	73	62				
5	4	76	74	69				
6	5	71	69	75				
7	6	85	82	67				
8								
9	Anova: Fator Único							
10								
11	RESUMO							
12	*Grupos*	*Contagem*	*Soma*	*Média*	*Variância*			
13	Atlanta	6	474	79	34			
14	Dallas	6	444	74	20			
15	Seattle	6	396	66	32			
16								
17								
18	ANOVA							
19	*Fonte de Variação*	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*	
20	Entre Grupos	516	2	258	9	0,0027	3,68	
21	Dentro de Grupos	430	15	28,6667				
22								
23	Total	946	17					
24								

Etapa 3. Quando a caixa de diálogo Análise de Dados aparecer:
Escolha a opção **Anova: Fator único** na lista Ferramentas de Análise
Dê um clique em **OK**

Etapa 4. Quando a caixa de diálogo Anova: Fator Único aparecer:
Digite B1:D7 na caixa **Intervalo de Entrada**
Marque a opção **Colunas**
Marque a opção **Rótulos na primeira linha**
Marque a opção **Intervalo de saída**
Dê um clique em **OK**

Comparações Envolvendo Proporções e Teste de Independência

ESTATÍSTICA NA PRÁTICA

UNITED WAY*
Rochester, Nova York

A United Way of Greater Rochester é uma organização sem fins lucrativos dedicada a melhorar a qualidade de vida das pessoas dos sete municípios aos quais serve, suprindo as necessidades mais importantes de assistência humana à comunidade.

A campanha anual de arrecadação de fundos da United Way/Red Cross, realizada a cada primavera, financia centenas de programas empreendidos por mais de 200 fornecedores de serviços. Esses fornecedores atendem a ampla variedade de necessidades humanas – físicas, intelectuais e sociais – e atendem a pessoas de todas as idades, origens e níveis econômicos.

Graças ao enorme envolvimento de voluntários, a United Way of Greater Rochester é capaz de manter seus custos operacionais em apenas oito centavos de cada dólar arrecadado.

A United Way of Great Rochester decidiu realizar uma pesquisa para conhecer melhor qual é a percepção que a comunidade tem de suas obras assistenciais. Entrevistas com grupos de foco (*focus group*) foram

* Os autores agradecem ao Dr. Philip R. Tyler, Consultor de Marketing da United Way, por fornecer esta "Estatística na Prática".

realizadas com profissionais, prestadores de serviços e trabalhadores em geral para obter informações preliminares sobre a percepção das pessoas a respeito de seu trabalho. As informações obtidas foram então usadas para ajudar a desenvolver o questionário da pesquisa. O questionário foi testado previamente, modificado e distribuído a 440 pessoas; 323 questionários preenchidos foram devolvidos.

Uma série de estatísticas descritivas, incluindo distribuições de freqüência e tabulações cruzadas, foi obtida dos dados coletados. Uma parte importante da análise envolveu o uso de tabelas de contingência e testes de independência quiquadrado. Uma função desses testes estatísticos era determinar se a percepção que as pessoas tinham dos gastos administrativos independia da ocupação por elas exercida.

As hipóteses do teste de independência eram:

H_0: A percepção dos gastos administrativos da United Way independe da profissão do entrevistado.

H_a: A percepção dos gastos administrativos da United depende da profissão do entrevistado.

Duas perguntas da pesquisa forneceram os dados para o teste estatístico. Uma das perguntas obteve dados sobre a percepção que as pessoas tinham da porcentagem dos fundos arrecadados destinada a despesas administrativas (até 10%, de 11% a 20% e 21% ou mais). A outra questão perguntava a profissão do entrevistado.

O teste do quiquadrado ao nível de significância de 0,05 levou à rejeição da hipótese nula de independência e à conclusão de que a percepção que as pessoas tinham dos gastos administrativos da United Way variava de acordo com a profissão. As despesas administrativas reais eram inferiores a 9%, mas 35% dos entrevistados achavam que as despesas administrativas eram de 21% ou mais. Portanto, muitos tinham percepções equivocadas dos custos administrativos. Nesse grupo, empregados de linhas de produção, funcionários de escritórios, de equipes de vendas e da área técnica e profissional apresentaram percepções mais equivocadas que os dos demais grupos.

O estudo da percepção existente entre a comunidade ajudou a United Way of Rochester a desenvolver ajustes ao seu programa e às suas atividades de arrecadação de fundos. Neste capítulo, você aprenderá como um teste estatístico de independência, similar ao que acabamos de descrever, é realizado.

Muitas aplicações estatísticas requerem uma comparação das proporções populacionais. Na Seção 11.1, descreveremos inferências estatísticas com respeito às diferenças entre as proporções de duas populações. Duas amostras são necessárias, sendo uma de cada população, e a inferência estatística baseia-se nas duas proporções amostrais. A segunda seção examina um teste de hipóteses que compara as proporções de uma única população multinomial com as proporções estabelecidas em uma hipótese nula. Uma amostra da população multinomial é usada, e o teste de hipóteses baseia-se em comparar as proporções amostrais com as que foram estabelecidas na hipótese nula. Na última seção do capítulo, vamos mostrar como as tabelas de contingência podem ser usadas para testar a independência de duas variáveis. Uma amostra é usada para o teste de independência, mas medidas das duas variáveis são necessárias para cada elemento amostrado. Ambas as Seções 11.2 e 11.3 recorrem ao uso de um teste estatístico quiquadrado.

11.1 INFERÊNCIAS SOBRE A DIFERENÇA ENTRE AS PROPORÇÕES DE DUAS POPULAÇÕES

Admitindo que p_1 denota a proporção da população 1 e que p_2 denota a proporção da população 2, consideramos inferências sobre a diferença entre as proporções de duas populações: $p_1 - p_2$. Para fazer uma inferência sobre essa diferença, selecionaremos duas amostras aleatórias que consistem em n_1 unidades da população 1 e n_2 unidades da população 2.

Estimação por Intervalo de $p_1 - p_2$

No exemplo seguinte, mostramos como calcular uma margem de erro e desenvolver uma estimação por intervalo da diferença entre duas proporções populacionais.

Uma firma especializada em declarações do imposto de renda está interessada em comparar a qualidade do trabalho em dois de seus escritórios regionais. Ao selecionar aleatoriamente amostras de declarações do imposto de renda preenchidas em cada escritório e verificar a precisão amostral das declarações, a firma

será capaz de estimar a proporção das declarações preenchidas erroneamente em cada escritório. Interessa-lhes especialmente a diferença entre essas proporções.

p_1 = proporção de declarações preenchidas erroneamente para a população 1 (escritório 1).
p_2 = proporção de declarações preenchidas erroneamente para a população 2 (escritório 2).
$\bar{p}_1$ = proporção amostral de uma amostra aleatória simples extraída da população 1.
$\bar{p}_2$ = proporção amostral de uma amostra aleatória simples extraída da população 2.

A diferença entre as duas proporções amostrais é dada por $p_1 - p_2$. O estimador por ponto de $p_1 - p_2$ é o seguinte:

ESTIMADOR POR PONTO DA DIFERENÇA ENTRE DUAS PROPORÇÕES POPULACIONAIS

$$\bar{p}_1 - \bar{p}_2 \qquad \textbf{(11.1)}$$

Desse modo, o estimador por ponto da diferença entre duas proporções populacionais é a diferença entre as proporções amostrais de duas amostras aleatórias simples independentes.

À semelhança do que ocorre com outros estimadores por ponto, o estimador por ponto $\bar{p}_1 - \bar{p}_2$ tem uma distribuição amostral que reflete os valores possíveis de $\bar{p}_1 - \bar{p}_2$ se tomássemos, repetidamente, duas amostras aleatórias independentes. A média desta distribuição amostral é $p_1 - p_2$ e o erro padrão de $\bar{p}_1 - \bar{p}_2$ é o seguinte:

$$\sigma_{\bar{p}_1-\bar{p}_2} = \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}} \qquad \textbf{(11.2)}$$

Se os tamanhos de amostra forem suficientemente grandes a ponto de n_1p_1, $n_1(1-p_1)$, n_2p_2 e $n_2(1-p_2)$ serem todos maiores ou iguais a 5, a distribuição amostral de $\bar{p}_1 - \bar{p}_2$ pode ser aproximada a uma distribuição normal.

Conforme mostramos anteriormente, uma estimação por intervalo é dada por uma estimação por ponto ± uma margem de erro. Na estimação da diferença entre duas proporções populacionais, uma estimação por intervalo assumirá a seguinte forma:

$$\bar{p}_1 - \bar{p}_2 \pm \text{Margem de erro}$$

Com a distribuição amostral de $\bar{p}_1 - \bar{p}_2$ aproximada a uma distribuição normal, desejaríamos usar $z_{\alpha/2}\, \sigma_{\bar{p}_1-\bar{p}_2}$ como a margem de erro. Entretanto, o $\sigma_{\bar{p}_1-\bar{p}_2}$ dado pela Equação 11.2 não pode ser usado diretamente porque as duas proporções populacionais, p_1 e p_2 são desconhecidas. Usando a proporção amostral $\bar{p}_1$ para estimar p_1 e a proporção amostral $\bar{p}_2$ para estimar p_2, a margem de erro é a seguinte:

$$\textit{Margem de erro} = z_{\alpha/2}\sqrt{\frac{\bar{p}_1(1-\bar{p}_1)}{n_1} + \frac{\bar{p}_2(1-\bar{p}_2)}{n_2}} \qquad \textbf{(11.3)}$$

A forma geral de uma estimação por intervalo entre duas proporções populacionais é a seguinte:

ESTIMAÇÃO POR INTERVALO DA DIFERENÇA ENTRE DUAS PROPORÇÕES POPULACIONAIS

$$\bar{p}_1 - \bar{p}_2 \pm z_{\alpha/2}\sqrt{\frac{\bar{p}_1(1-\bar{p}_1)}{n_1} + \frac{\bar{p}_2(1-\bar{p}_2)}{n_2}} \qquad \textbf{(11.4)}$$

em que $1 - \alpha$ é o coeficiente de confiança.

Retornando ao exemplo das declarações de imposto de renda, descobrimos que as amostras aleatórias simples independentes dos dois escritórios fornecem as seguintes informações:

Escritório 1	Escritório 2
$n_1 = 250$	$n_2 = 300$
Número de declarações com erros = 35	Número de declarações com erros = 27

ARQUIVO DA INTERNET
TaxPrep

As proporções amostrais correspondentes aos dois escritórios são as seguintes:

$$\bar{p}_1 = \frac{35}{250} = 0{,}14$$

$$\bar{p}_2 = \frac{27}{300} = 0{,}09$$

A estimação por ponto da diferença entre as proporções de declarações errôneas do imposto de renda para as duas populações é $\bar{p}_1 - \bar{p}_2 = 0{,}14 - 0{,}09 = 0{,}05$. Desse modo, estimamos que o escritório 1 tem um índice de erro igual a 0,05, ou 5%, maior que a do escritório 2.

A Equação 11.4 agora pode ser usada para fornecer uma margem de erro e a estimação por intervalo da diferença entre as duas proporções populacionais. Usando um intervalo de confiança de 90% com $z_{\alpha/2} = z_{0,05} = 1{,}645$, obtemos:

$$\bar{p}_1 - \bar{p}_2 \pm z_{\alpha/2}\sqrt{\frac{\bar{p}_1(1-\bar{p}_1)}{n_1} + \frac{\bar{p}_2(1-\bar{p}_2)}{n_2}}$$

$$0{,}14 - 0{,}09 \pm 1{,}645\sqrt{\frac{0{,}14(1-0{,}14)}{250} + \frac{0{,}09(1-0{,}09)}{300}}$$

$$0{,}05 \pm 0{,}045$$

Assim, a margem de erro é 0,045, e o intervalo de confiança de 90% varia de 0,005 a 0,095.

Testes de Hipóteses sobre $p_1 - p_2$

Todas as hipóteses consideradas utilizam 0 como a diferença de interesse.

Consideremos agora os testes de hipóteses sobre a diferença entre as proporções de duas populações. Vamos nos concentrar em testes que não envolvem diferenças entre as duas proporções populacionais. Nesse caso, as três formas de teste de hipóteses são as seguintes:

$$H_0: p_1 - p_2 \geq 0 \qquad H_0: p_1 - p_2 \leq 0 \qquad H_0: p_1 - p_2 = 0$$
$$H_a: p_1 - p_2 < 0 \qquad H_a: p_1 - p_2 > 0 \qquad H_a: p_1 - p_2 \neq 0$$

Quando assumimos que H_0 é verdadeira enquanto igualdade, temos $p_1 - p_2 = 0$, que equivale a dizer que as proporções populacionais são iguais, $p_1 = p_2$.

Basearemos a estatística de teste na distribuição amostral do estimador por ponto $\bar{p}_1 - \bar{p}_2$. Na Equação 11.2, mostramos que o erro padrão de $\bar{p}_1 - \bar{p}_2$ é dado por:

$$\sigma_{\bar{p}_1 - \bar{p}_2} = \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$$

Sob a hipótese de que H_0 é verdadeira enquanto igualdade, as proporções populacionais são iguais e $p_1 - p_2 = p$. Nesse caso, $\sigma_{\bar{p}_1 - \bar{p}_2}$ torna-se:

$$\sigma_{\bar{p}_1 - \bar{p}_2} = \sqrt{\frac{p(1-p)}{n_1} + \frac{p(1-p)}{n_2}} = \sqrt{p(1-p)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)} \qquad \textbf{(11.5)}$$

Com p desconhecido, agrupamos, ou combinamos, os estimadores por ponto das duas amostras ($\bar{p}_1$ e $\bar{p}_2$) para obtermos um único estimador por ponto de p da seguinte maneira:

$$\bar{p} = \frac{n_1\bar{p}_1 + n_2\bar{p}_2}{n_1 + n_2} \qquad \textbf{(11.6)}$$

Esse **estimador agrupado** (*pooled estimator*) de p é uma média ponderada de $\bar{p}_1$ e $\bar{p}_2$.

Substituindo $\bar{p}$ por p na Equação 11.5, obtemos uma estimativa do erro padrão de $\bar{p}_1 - \bar{p}_2$. Essa estimativa do erro padrão é usada na estatística de teste. A forma geral da estatística de teste para testes de hipóteses sobre a diferença entre duas proporções populacionais é o estimador por ponto dividido pela estimativa de $\sigma_{\bar{p}_1 - \bar{p}_2}$.

ESTÁTISTICA DE TESTE PARA TESTES DE HIPÓTESES SOBRE $p_1 - p_2$

$$z = \frac{(\bar{p}_1 - \bar{p}_2)}{\sqrt{\bar{p}(1-\bar{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \tag{11.7}$$

Essa estatística de teste aplica-se a situações com grandes amostras, em que n_1p_1, $n_1(1 - p_1)$, n_2p_2 e n_2 $(1 - p_2)$ são todas maiores ou iguais a 5.

Retornemos ao exemplo da firma especializada em declarações do imposto de renda e vamos supor que a firma queira usar um teste de hipóteses para determinar se as proporções de erro diferem entre os dois escritórios. Nesse caso, é necessário um teste bicaudal. As hipóteses nula e alternativa são as seguintes:

$$H_0: p_1 - p_2 = 0$$
$$H_a: p_1 - p_2 \neq 0$$

Se H_0 for rejeitada, a firma poderá concluir que os índices de erro nos dois escritórios diferem. Usaremos $\alpha = 0{,}10$ como nível de significância.

Os dados amostrais coletados anteriormente mostraram que $\bar{p}_1 = 0{,}14$ para as $n_1 = 250$ declarações amostradas no escritório 1 e $\bar{p}_2 = 0{,}09$ para as $n_2 = 300$ declarações amostradas no escritório 2. Prosseguimos os cálculos da estimativa agrupada de p:

$$\bar{p} = \frac{n_1\bar{p}_1 + n_2\bar{p}_2}{n_1 + n_2} = \frac{250(0{,}14) + 300(0{,}09)}{250 + 300} = 0{,}1127$$

Usando essa estimativa agrupada e a diferença entre as proporções amostrais, o valor da estatística de teste é o seguinte:

$$z = \frac{(\bar{p}_1 - \bar{p}_2)}{\sqrt{\bar{p}(1-\bar{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} = \frac{(0{,}14 - 0{,}09)}{\sqrt{0{,}1127(1 - 0{,}1127)\left(\frac{1}{250} + \frac{1}{300}\right)}} = 1{,}85$$

Ao calcularmos o valor p para esse teste bicaudal, observamos primeiramente que $z = 1{,}85$ na cauda superior da distribuição normal padrão. Usando $z = 1{,}85$ e a tabela de distribuição normal padrão, descobrimos que a área na cauda superior é $0{,}5000 - 0{,}4678 = 0{,}0322$. Duplicando essa área para um teste bicaudal, encontramos o valor $p = 2(0{,}0322) = 0{,}0644$. Com o valor p menor que $\alpha = 0{,}10$, H_0 é rejeitada ao nível de significância 0,10. A firma pode concluir que os índices de erro diferem entre os dois escritórios. Essa conclusão do teste de hipóteses é coerente com os resultados anteriores de estimação por intervalo que mostraram que a estimação por intervalo da diferença entre os índices de erro populacionais nos dois escritórios variam de 0,005 a 0,095, sendo que o escritório 1 apresenta o maior índice de erros.

Exercícios

Métodos

1. Considere os seguintes resultados de amostras independentes tomadas de duas populações:

Amostra 1	Amostra 2
$n_1 = 400$	$n_2 = 300$
$\bar{p}_1 = 0{,}48$	$\bar{p}_2 = 0{,}36$

a. Qual é a estimação por ponto da diferença entre as duas proporções populacionais?

b. Desenvolva um intervalo de confiança de 90% para a diferença entre as duas proporções populacionais.

c. Desenvolva um intervalo de confiança de 95% para a diferença entre as duas proporções populacionais.

2. Considere o teste de hipóteses:

$$H_0: p_1 - p_2 \leq 0$$
$$H_a: p_1 - p_2 > 0$$

Os resultados a seguir referem-se a amostras independentes tomadas de duas populações:

Amostra 1	Amostra 2
$n_1 = 200$	$n_2 = 300$
$\bar{p}_1 = 0{,}22$	$\bar{p}_2 = 0{,}16$

a. Qual é o valor p?
b. Com $\alpha = 0{,}05$, qual é a conclusão do seu teste de hipóteses?

Aplicações

3. Uma pesquisa realizada pela *Business Week*/Harris perguntou a altos executivos de grandes corporações quais eram suas opiniões a respeito do panorama econômico para o futuro. Uma das perguntas foi a seguinte: "Você acha que haverá um aumento no número de empregados em tempo integral em sua empresa nos próximos 12 meses?" Na pesquisa atual, 220 de 400 executivos responderam sim, ao passo que, na pesquisa realizada no ano anterior, 192 de 400 executivos responderam sim. Forneça uma estimação por intervalo de confiança de 95% da diferença entre as proporções nos dois períodos. Qual é a sua interpretação da estimação por intervalo?

4. Nos últimos anos, o número de pessoas que usam a internet para obter notícias políticas aumentou. Freqüentemente, os sites da Web de partidos políticos pedem aos internautas para registrarem suas opiniões em pesquisas on-line. O Pew Research Center realizou uma pesquisa própria para saber qual era a participação de republicanos e democratas nas pesquisas on-line (Associated Press, 6 de janeiro de 2003). Aplicam-se os seguintes dados amostrais.

Partido Político	Tamanho da Amostra	Participam de Pesquisas On-Line
Republicano	250	115
Democrata	350	98

a. Calcule a estimação por ponto da proporção de republicanos que indicam que participariam de pesquisas on-line. Calcule a estimação por ponto relativa aos democratas.
b. Qual é a estimação por ponto da diferença entre as duas proporções populacionais?
c. Com 95% de confiança, qual é a margem de erro?
d. Representantes das instituições de pesquisa científica afirmam que a profusão de pesquisas on-line pode confundir as pessoas a respeito da opinião pública real. Você concorda com essa afirmação? Use uma estimação por intervalo de confiança de 95% da diferença entre as proporções populacionais de republicanos e democratas para ajudar a justificar sua resposta.

5. Os caça-níqueis são o jogo predileto nos cassinos de todo o território nacional nos Estados Unidos (*Harrah's Survey 2002: Profile of the American Gamble*). Os seguintes dados amostrais exibem o número de mulheres e de homens que escolheram os caça-níqueis como o jogo favorito.

	Mulheres	Homens
Tamanho da amostra	320	250
Jogo favorito: caça-níqueis	256	165

a. Qual é a estimação por ponto da proporção de mulheres que dizem que os caça-níqueis são seu jogo favorito?
b. Qual é a estimação por ponto da proporção de homens que dizem que os caça-níqueis são seu jogo favorito?
c. Forneça uma estimação por intervalo de confiança de 95% da diferença entre a proporção de mulheres e da proporção de homens que dizem que os caça-níqueis são o jogo favorito.

6. O Bureau of Transportation faz um acompanhamento do desempenho das dez maiores empresas aéreas dos Estados Unidos quanto aos horários de chegada de seus vôos (*The Wall Street Journal*, 4 de março de 2003). Os vôos que chegam em um intervalo de 15 minutos do horário programado são considerados pontuais. Usando dados amostrais coerentes com as estatísticas do Bureau of Transportation publicadas em janeiro de 2001 e em janeiro de 2002, considere o seguinte:

Janeiro de 2001 Uma amostra de 924 vôos apresentou 742 que chegaram no horário.
Janeiro de 2002 Uma amostra de 841 vôos apresentou 714 que chegaram no horário.

a. Qual é a estimação por ponto dos vôos que chegaram no horário em janeiro de 2001?
b. Qual é a estimação por ponto dos vôos que chegaram no horário em janeiro de 2002?
c. Digamos que p_1 denote a proporção populacional dos vôos que chegaram no horário em janeiro de 2001 e que p_2 denote a proporção populacional dos vôos que chegaram no horário em janeiro de 2002. Estabeleça as hipóteses que poderiam ser testadas para determinar se as principais empresas aéreas melhoraram seu desempenho quanto à chegada de vôos durante o período de um ano.
d. Qual é o valor p? Com $\alpha = 0{,}05$, qual é a sua conclusão?

7. Em um teste da qualidade de dois comerciais de televisão, cada comercial foi exibido seis vezes em regiões de teste distintas durante o período de uma semana. Uma semana depois, foi realizada uma pesquisa telefônica para identificar as pessoas que assistiram aos comerciais. Essas pessoas foram solicitadas a dizer qual foi a mensagem principal dos comerciais. Foram registrados os seguintes resultados:

	Comercial A	Comercial B
Número de Pessoas que Assistiram ao Comercial	150	200
Número de Pessoas que se Lembravam da Mensagem	63	60

a. Use $\alpha = 0{,}05$ e teste a hipótese de não haver nenhuma diferença nas proporções de lembrança referentes aos dois comerciais.
b. Calcule um intervalo de confiança de 95% para a diferença entre as proporções de lembrança para as duas populações.

8. Durante o Super Bowl de 2003, o comercial da Miller Lite Beer, chamado "The Miller Lite Girls", classificou-se entre os três anúncios mais eficazes veiculados durante o Super Bowl (*USA Today*, 29 de dezembro de 2003). A avaliação da eficácia publicitária, realizada pela pesquisa Ad Track do jornal *USA Today*, divulgou amostras separadas de acordo com a faixa etária dos entrevistados para saber como os anúncios veiculados durante o Super Bowl chamavam a atenção dos diferentes grupos etários. Os dados amostrais seguintes aplicam-se ao comercial "The Miller Lite Girls".

Faixa etária	Tamanho da amostra	Gostaram muito do anúncio
Menos de 30 anos	100	49
De 30 a 49 anos	150	54

a. Formule um teste de hipóteses que possa ser usado para determinar se há uma diferença entre as proporções populacionais correspondentes aos dois grupos etários.
b. Qual é a estimação por ponto da diferença entre as duas proporções populacionais?
c. Realize um teste de hipóteses e relate o valor p. Com $\alpha = 0{,}05$, qual é a sua conclusão?
d. Discuta o atrativo dos anúncios para os grupos etários dos mais jovens e dos mais velhos. A organização Miller Lite consideraria encorajadores os resultados obtidos pela pesquisa Ad Track do jornal *USA Today*? Explique.

9. Uma pesquisa de opinião do *New York Times*/CBS News realizada em 2003 tomou como amostra 523 adultos que planejavam férias para os próximos seis meses e descobriu que 141 esperavam viajar de avião (*New York Times News Service*, 2 de março de 2003). Uma pesquisa idêntica realizada em maio de 1993 pelo *New York Times*/CBS descobriu que dos 477 adultos que planejavam férias para os próximos seis meses, 81 esperavam viajar de avião.

a. Estabeleça a hipótese que pode ser usada para determinar se ocorreu uma mudança significativa na proporção da população que planeja viajar de avião no período de dez anos.
b. Qual é a proporção amostral que espera viajar de avião em 2003? E em 1993?
c. Use $\alpha = 0{,}01$ e teste se há alguma diferença significativa. Qual é a sua conclusão?
d. Discuta as razões que poderiam fornecer uma explicação para essa conclusão.

10. A revista *Yahoo! Internet Life* patrocinou pesquisas em diversas regiões metropolitanas para estimar a proporção de adultos que usam a internet no trabalho (*USA Today*, 7 de maio de 2000). Os resultados revelaram que 40% dos adultos que moram em Washington, D.C., usam a internet no trabalho, enquanto 32% dos adultos de São Francisco usam a internet no trabalho. Se os tamanhos de amostra são 240 e 250, respectivamente, os resultados amostrais indicam que a proporção populacional de adultos que usam a internet no trabalho em Washington, D.C., é maior que a proporção populacional de São Francisco? Qual é o valor p? Usando $\alpha = 0{,}05$, qual é a conclusão?

11.2 TESTES DE HIPÓTESES PARA PROPORÇÕES DE UMA POPULAÇÃO MULTINOMIAL

Nesta seção, consideraremos os testes de hipóteses referentes à proporção de elementos de uma população pertencentes a cada uma das várias classes ou categorias. Diferentemente da seção anterior, lidaremos com uma única população: uma **população multinomial**. Os parâmetros da população multinomial são a proporção de elementos pertencentes a cada categoria; o teste de hipóteses que descrevemos refere-se ao valor desses parâmetros. A distribuição multinomial de probabilidade pode ser imaginada como uma extensão da distribuição binomial para o caso de três ou mais categorias de resultados. Em cada ensaio de um experimento multinomial, ocorre um e somente um dos resultados. Presume-se que cada ensaio do experimento seja independente, e as probabilidades dos resultados permanecem as mesmas para cada ensaio.

As hipóteses (suposições) referentes ao experimento multinomial têm uma estreita correspondência com as do experimento binomial, com exceção de que o experimento multinomial tem três ou mais resultados por ensaio.

Como exemplo, considere o estudo sobre participação no mercado realizado pela Scott Marketing Research. Durante o ano passado a participação no mercado permaneceu em 30% para a empresa A, 50% para a empresa B e 20% para a empresa C. Recentemente, a empresa C desenvolveu um "novo e melhorado" produto para substituir seu atual lançamento no mercado. A empresa C assinou um contrato de consultoria com a Scott Marketing Research para determinar se o novo produto alterará as fatias de mercado.

Nesse caso, a população de interesse é uma população multinomial; cada cliente é classificado como alguém que compra da empresa A, B ou C. Dessa forma, temos uma população multinomial com três resultados. Vamos usar a seguinte notação para as proporções:

$$p_A = \text{fatia de mercado da empresa A}$$
$$p_B = \text{fatia de mercado da empresa B}$$
$$p_C = \text{fatia de mercado da empresa C}$$

A Scott Marketing Research realizará uma pesquisa amostral e calculará a proporção dos que preferem o produto de cada uma das empresas. Um teste de hipóteses será realizado então para verificar se o novo produto causou alguma alteração nas fatias de mercado. Supondo que o novo produto da empresa C não altere as fatias de mercado, as hipóteses nula e alternativa são estabelecidas da seguinte maneira:

$$H_0: p_A = 0{,}30, p_B = 0{,}50 \text{ e } p_C = 0{,}20$$
$$H_a: \text{As proporções populacionais não são}$$
$$p_A = 0{,}30, p_B = 0{,}50 \text{ e } p_C = 0{,}20$$

Se os resultados amostrais levarem à rejeição de H_0, a Scott Marketing Research terá evidências de que a introdução do novo produto afeta as participações no mercado.

Vamos supor que a firma de pesquisa de mercado tenha usado no estudo um painel de consumo composto de 200 consumidores. Cada indivíduo foi solicitado a especificar uma preferência de compra entre as três alternativas: o produto da empresa A, o produto da empresa B e o novo produto da empresa C. As 200 respostas estão resumidas a seguir:

O painel de consumo composto de 200 consumidores no qual cada pessoa é solicitada a escolher uma de três alternativas é equivalente a um experimento multinomial composto de 200 ensaios.

	Freqüência Observada	
Produto da Empresa A	Produto da Empresa B	Novo Produto da Empresa C
48	98	54

Agora, podemos executar um **teste da eficiência de ajuste**, o qual determinará se a amostra das preferências de compra da parte de 200 consumidores é coerente com a hipótese nula. O teste de eficiência de ajuste baseia-se em uma comparação da amostra de resultados *observados* com os resultados *esperados* sob a suposição de que a hipótese nula é verdadeira. Portanto, o passo seguinte é calcular as preferências de compra esperadas dos 200 clientes sob a suposição de que $p_A = 0{,}30$, $p_B = 0{,}50$ e $p_C = 0{,}20$. Essa operação produz os resultados esperados.

	Freqüência Esperada	
Produto da Empresa A	Produto da Empresa B	Novo Produto da Empresa C
200(0,30) = 60	200(0,50) = 100	200(0,20) = 40

Assim, notamos que a freqüência esperada correspondente a cada categoria é encontrada multiplicando-se o tamanho da amostra, 200, pela proporção hipotética da categoria.

O teste de eficiência de ajuste agora se concentra nas diferenças entre as freqüências observadas e as freqüências esperadas. Diferenças grandes entre as freqüências observadas e as freqüências esperadas suscitam dúvidas sobre a suposição de que as proporções hipotéticas ou fatias de mercado estejam corretas. A questão referente a se as diferenças entre as freqüências observadas e esperadas são "grandes" ou "pequenas" é respondida com a ajuda da seguinte estatística de teste.

ESTATÍSTICA DE TESTE PARA A EFICIÊNCIA DE AJUSTE

$$\chi^2 = \sum_{i=1}^{k} \frac{(f_i - e_i)^2}{e_i} \qquad (11.8)$$

em que

f_i = freqüência observada para a categoria i
e_i = freqüência esperada para a categoria i
k = o número de categorias

Observação: A estatística de teste tem uma distribuição quiquadrado com $k - 1$ graus de liberdade, desde que as freqüências esperadas sejam 5 *ou mais* para todas as categorias.

O teste de eficiência de ajuste é sempre um teste unicaudal, e a rejeição ocorre na cauda superior da distribuição quiquadrado.

Continuemos com o exemplo da Scott Marketing Research, no qual utilizaremos os dados amostrais para testar a hipótese de que a população multinomial mantém as proporções $P_A = 0{,}30$, $P_B = 0{,}50$ e $P_C = 0{,}20$. Usaremos um nível de significância $\alpha = 0{,}05$. Prosseguimos, usando as freqüências observadas e esperadas para calcular o valor da estatística de teste. Sendo 5 ou mais todas as freqüências esperadas, o cálculo da estatística de teste quiquadrado é apresentado na Tabela 11.1. Desse modo, temos $\chi^2 = 7{,}34$.

Rejeitaremos a hipótese nula se as diferenças entre as freqüências observadas e esperadas forem *grandes*. Diferenças grandes entre as freqüências observadas e esperadas resultarão em um valor grande para a estatística de teste. Assim, o teste de eficiência de ajuste sempre será um teste da cauda superior. Usaremos a área da cauda superior de uma distribuição quiquadrado e o critério do valor p para determinar se a hipótese nula pode ser rejeitada. Com $k - 1 = 3 - 1 = 2$ graus de liberdade, a Tabela 11.2 exibe as seguintes áreas na cauda superior e seus valores quiquadrado (χ^2) correspondentes:

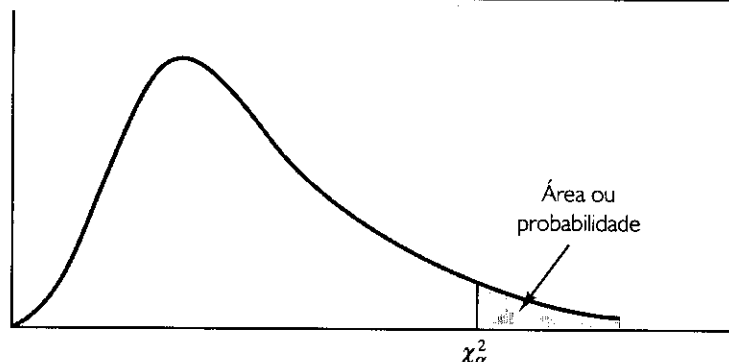
Área na Cauda Superior	0,10	0,05	0,025	0,01
Valor χ^2 (2 gl)	4,605	5,991	7,378	9,210

↑ $\chi^2 = 7{,}34$ (entre 5,991 e 7,378)

A estatística de teste $\chi^2 = 7{,}34$ está entre 5,991 e 7,378. Desse modo, a área da cauda superior, ou valor p, correspondente, deve estar entre 0,05 e 0,025. Com o valor $p \leq \alpha = 0{,}05$, rejeitamos H_0 e concluímos que a introdução no mercado do novo produto da empresa C alterará a atual estrutura de participação no mercado. O Minitab ou o Excel podem ser usados para demonstrar que $\chi^2 = 7{,}34$ produz um valor $p = 0{,}0255$.

Tabela 11.1 Cálculo da estatística de teste quiquadrado para o estudo de participação no mercado realizado pela Scott Marketing Research

Categoria	Proporção Hipotética	Freqüência Observada (f_i)	Freqüência Esperada (e_i)	Diferença ($f_i - e_i$)	Quadrado da Diferença $(f_i - e_i)^2$	Quadrado Dividido pela Freqüência Esperada $(f_i - e_i)^2/e_i$
Empresa A	0,30	48	60	−12	144	2,40
Empresa B	0,50	98	100	−2	4	0,04
Empresa C	0,20	54	40	14	196	4,90
Total		200				$\chi^2 = 7{,}34$

Tabela 11.2 Valores selecionados da tabela de distribuição quiquadrado*

Graus de Liberdade	Área da Cauda Superior							
	0,99	0,975	0,95	0,90	0,10	0,05	0,025	0,01
1	0,000	0,001	0,004	0,016	2,706	3,841	5,024	6,635
2	0,020	0,051	0,103	0,211	4,605	5,991	7,378	9,210
3	0,115	0,216	0,352	0,584	6,251	7,815	9,348	11,345
4	0,297	0,484	0,711	1,064	7,779	9,488	11,143	13,277
5	0,554	0,831	1,145	1,610	9,236	11,070	12,832	15,086
6	0,872	1,237	1,635	2,204	10,645	12,592	14,449	16,812
7	1,239	1,690	2,167	2,833	12,017	14,067	16,013	18,475
8	1,647	2,180	2,733	3,490	13,362	15,507	17,535	20,090
9	2,088	2,700	3,325	4,168	14,684	16,919	19,023	21,666
10	2,558	3,247	3,940	4,865	15,987	18,307	20,483	23,209
11	3,053	3,816	4,575	5,578	17,275	19,675	21,920	24,725
12	3,571	4,404	5,226	6,304	18,549	21,026	23,337	26,217
13	4,107	5,009	5,892	7,041	19,812	22,362	24,736	27,688
14	4,660	5,629	6,571	7,790	21,064	23,685	26,119	29,141
15	5,229	6,262	7,261	8,547	22,307	24,996	27,488	30,578
16	5,812	6,908	7,962	9,312	23,542	26,296	28,845	32,000
17	6,408	7,564	8,672	10,085	24,769	27,587	30,191	33,409
18	7,015	8,231	9,390	10,865	25,989	28,869	31,526	34,805
19	7,633	8,907	10,117	11,651	27,204	30,144	32,852	36,191
20	8,260	9,591	10,851	12,443	28,412	31,410	34,170	37,566
21	8,897	10,283	11,591	13,240	29,615	32,671	35,479	38,932
22	9,542	10,982	12,338	14,041	30,813	33,924	36,781	40,289
23	10,196	11,689	13,091	14,848	32,007	35,172	38,076	41,638
24	10,856	12,401	13,848	15,659	33,196	36,415	39,364	42,980
25	11,524	13,120	14,611	16,473	34,382	37,652	40,646	44,314
26	12,198	13,844	15,379	17,292	35,563	38,885	41,923	45,642
27	12,878	14,573	16,151	18,114	36,741	40,113	43,195	46,963
28	13,565	15,308	16,928	18,939	37,916	41,337	44,461	48,278
29	14,256	16,047	17,708	19,768	39,087	42,557	45,722	49,588
30	14,953	16,791	18,493	20,599	40,256	43,773	46,979	50,892
40	22,164	24,433	26,509	29,051	51,805	55,758	59,342	63,691
60	37,485	40,482	43,188	46,459	74,397	79,082	83,298	88,379
80	53,540	57,153	60,391	64,278	96,578	101,879	106,629	112,329
100	70,065	74,222	77,929	82,358	118,498	124,342	129,561	135,807

*Nota: A Tabela 3 do Apêndice B é a mais completa.

Em vez de usar o valor p, poderíamos utilizar o critério do valor crítico para chegar à mesma conclusão. Com $\alpha = 0{,}05$ e 2 graus de liberdade, o valor crítico da estatística de teste é $\chi^2_{0,05} = 5{,}991$. A regra de rejeição para a cauda superior torna-se

$$\text{Rejeitar } H_0 \text{ se } \chi^2 \geq 5{,}991$$

Com 7,34 > 5,991, rejeitamos H_0. O critério do valor p e o critério do valor crítico fornecem a mesma conclusão de teste de hipóteses.

Embora os resultados do teste não nos permitam tirar conclusões adicionais, podemos comparar informalmente as freqüências observadas e esperadas para obtermos uma idéia de como a estrutura de participação no mercado pode se alterar. Considerando a empresa C, descobrimos que a freqüência observada, 54, é maior que a freqüência esperada, 40. Visto que a freqüência esperada se baseou nas fatias de mercado atuais, a maior freqüência observada sugere que o novo produto terá um efeito positivo sobre a fatia de mercado da empresa C. Comparações entre as freqüências observada e esperada correspondentes às duas outras empresas indicam que o ganho de participação no mercado por parte da empresa C será mais prejudicial à empresa A do que à empresa B.

Vamos resumir as etapas gerais que podem ser usadas para se realizar um teste de eficiência de ajuste para uma distribuição populacional multinomial hipotética.

TESTE DE EFICIÊNCIA DE AJUSTE DA DISTRIBUIÇÃO MULTINOMIAL: RESUMO

1. Estabeleça as hipóteses nula e alternativa.

   H_0: A população segue uma distribuição multinomial com probabilidades específicas para cada uma das k categorias.

   H_a: A população não segue uma distribuição multinomial com as probabilidades especificadas para cada uma das k categorias.

2. Selecione uma amostra aleatória e registre as freqüências observadas f_i para cada categoria.
3. Suponha que a hipótese nula seja verdadeira e determine a freqüência esperada e_i em cada categoria multiplicando a probabilidade da categoria pelo tamanho da amostra.
4. Calcule o valor da estatística de teste

$$\chi^2 = \sum_{i=1}^{k} \frac{(f_i - e_i)^2}{e_i}$$

5. Regra de rejeição:

Critério do valor p:	Rejeitar H_0 se o valor $p \leq \alpha$
Critério do valor crítico:	Rejeitar H_0 se $\chi^2 \geq \chi^2_\alpha$

em que α é o nível de significância do teste e há $k - 1$ graus de liberdade.

Exercícios

Métodos

AUTOTESTE

11. Teste as seguintes hipóteses usando o teste de eficiência de ajuste χ^2.

$$H_0: p_A = 0{,}40,\ p_B = 0{,}40 \text{ e } p_C = 0{,}20$$
$$H_a: \text{As proporções populacionais não são } p_A = 0{,}40,\ p_B = 0{,}40 \text{ e } p_C = 0{,}20$$

Uma amostra de tamanho 200 produziu 60 na categoria A, 120 na categoria B e 20 na categoria C. Use $\alpha = 0{,}01$ e teste se as proporções estão em conformidade com o que foi declarado em H_0.

   a. Use o critério do valor p.
   b. Repita o teste usando o critério do valor crítico.

12. Suponha que temos uma população multinomial com quatro categorias A, B, C e D. A hipótese nula é que a proporção de itens é a mesma em cada categoria. A hipótese nula é:

$$H_0: p_A = p_B = p_C = p_D = 0{,}25$$

Um tamanho de amostra 300 produziu os seguintes resultados:

A: 85 B: 95 C: 50 D: 70

Use $\alpha = 0{,}05$ para determinar se H_0 deve ser rejeitada. Qual é o valor p?

Aplicações

AUTOTESTE

13. Durante as 13 primeiras semanas de uma nova série de televisão norte-americana, as proporções de público nos sábados à noite no horário das 8 às 9h foram registradas como 29% para a ABC, 28% para a CBS, 25% para a NBC e 18% para as emissoras independentes. Duas semanas depois de uma revisão da programação de sábado à noite, uma amostra de 300 residências produziu os seguintes dados de audiência: 95 sintonizavam a ABC; 70, a CBS; 89, a NBC; e 46, as emissoras independentes. Teste com $\alpha = 0{,}05$ para determinar se as proporções de público se alteraram.

14. A M&M/MARS, fabricantes dos M&M® Chocolate Candies (Confeitos de Chocolate M&M), realizou uma pesquisa nacional na qual os consumidores indicaram quais eram suas cores preferidas. Na brochura intitulada "Colors" que a M&M/MARS Consumer Affairs colocou à disposição, a distribuição tradicional de cores para os confeitos de chocolate é a seguinte:

Marrom	Amarelo	Vermelho	Laranja	Verde	Azul
30%	20%	20%	10%	10%	10%

Em um estudo de acompanhamento, pacotes de 453 g foram utilizados para determinar se as porcentagens registradas eram válidas. Os resultados seguintes foram obtidos para uma amostra de 506 confeitos de chocolate.

Marrom	Amarelo	Vermelho	Laranja	Verde	Azul
177	135	79	41	36	38

Use $\alpha = 0{,}05$ para determinar se esses dados confirmam as porcentagens divulgadas pela empresa.

15. Onde as mulheres compram roupas para o dia-a-dia mais freqüentemente? Dados do U.S. Shopper Database forneceram as seguintes porcentagens relativas às compras feitas por mulheres em cada uma das várias lojas (*The Wall Street Journal*, 28 de janeiro de 2003).

Loja	Porcentagem	Loja	Porcentagem
Wal-Mart	24%	Kohl's	8%
Lojas de departamento tradicionais	11%	Encomenda postal	12%
J.C. Penney	8%	Outros	37%

A outra categoria incluía lojas como a Target, a Kmart e a Sears, além de inúmeras lojas menores de produtos especiais. Nenhuma loja individual desse grupo envolvia mais de 5% das compradoras. Uma pesquisa recente usando uma amostra de 140 compradoras de Atlanta, Geórgia, revelou que 42 compravam no Wal-Mart; 20, em lojas de departamento tradicionais; 8, na J.C. Penney; 10, na Kohl's; 21, em lojas de encomenda postal; e 39, de outras fontes. Essa amostra sugere que as compradoras residentes em Atlanta diferem quanto às preferências de compra expressas no U.S. Shopper Database? Qual é o valor p? Use $\alpha = 0{,}05$. Qual é a sua conclusão?

16. A American Bankers Association coleta dados sobre o uso de cartões de crédito, cartões de débito, cheques pessoais e dinheiro vivo quando os consumidores pagam suas compras em lojas (*The Wall Street Journal*, 16 de dezembro de 2003). Em 1999, foram registrados os seguintes métodos de pagamento:

Compras na Loja	Porcentagem
Cartão de crédito	22%
Cartão de débito	21%
Cheque pessoal	18%
Dinheiro	39%

Uma amostra tomada em 2003 revelou que das 220 compras feitas em lojas, 46 usaram cartões de crédito; 67, cartões de débito; 33, cheques pessoais; e 74, dinheiro.

a. Com $\alpha = 0{,}01$, podemos concluir que ocorreu uma alteração na maneira pela qual os clientes pagavam as contas nas lojas no decorrer do período de quatro anos, de 1999 a 2003? Qual é o valor p?
b. Calcule a porcentagem de uso de cada método de pagamento usando os dados amostrais de 2003. Qual parece ter sido a maior mudança ou mudanças ao longo do período de quatro anos?
c. Em 2003, qual porcentagem de pagamentos foi feita com "dinheiro de plástico" (cartões de crédito ou cartões de débito)?

17. O Shareholder Scoreboard do *The Wall Street Journal* acompanha o desempenho de mil grandes empresas norte-americanas (*The Wall Street Journal*, 10 de março de 2003). O desempenho de cada empresa é classificado em função do retorno anual total, incluindo as variações nos preços das ações e no reinvestimento de dividendos. As classificações são atribuídas dividindo-se todas as mil empresas em grupos A (as 20% maiores), B (as 20% médias) a E (as 20% de nível mais baixo). Apresentamos a seguir as classificações no período de um ano correspondentes a uma amostra de 60 das maiores empresas. As maiores empresas diferem quanto ao desempenho das mil empresas integrantes do Shareholder Scoreboard? Use $\alpha = 0{,}05$.

A	B	C	D	E
5	8	15	20	12

18. Qual é a qualidades do serviço que as empresas aéreas prestam aos seus clientes? Um estudo revelou as seguintes avaliações dos clientes: 3% o consideraram excelente, 28% bom, 45% razoável e 24% ruim (*Business Week*, 11 de setembro de 2000). Em um estudo de acompanhamento dos serviços prestados pelas empresas, feito por telefone, suponha que uma amostra de 400 adultos tenha revelado as seguintes avaliações dos clientes: 24 consideraram o serviço excelente, 124 bom, 172 razoável e 80 ruim. A distribuição das avaliações dos clientes feitas por telefone a respeito das empresas é diferente da distribuição de avaliações dos clientes para as empresas aéreas? Teste com $\alpha = 0{,}01$. Qual é a sua conclusão?

11.3 TESTE DE INDEPENDÊNCIA

Outra aplicação importante da distribuição quiquadrado envolve usar os dados amostrais para testar a independência de duas variáveis. Vamos ilustrar o teste de independência considerando o estudo realizado pela Alber's Brewery of Tucson, no Arizona. A Alber's produz e distribui três tipos de cerveja: *light*, comum e escura.

Tabela 11.3 Tabela de contingência referente à preferência por um tipo de cerveja e ao sexo do consumidor

		Cerveja preferida		
		Light	Comum	Escura
Sexo	Masculino	célula(1,1)	célula(1,2)	célula(1,3)
	Feminino	célula(2,1)	célula(2,2)	célula(2,3)

Em uma análise dos segmentos de mercado das três cervejas, a equipe de pesquisa de mercado da empresa levantou a seguinte questão: As preferências pelos três tipos de cerveja diferem entre os consumidores masculinos e femininos? Se a preferência pela cerveja independer do sexo do consumidor, será iniciada uma campanha publicitária dirigida a todos os consumidores de cerveja. Entretanto, se a preferência pelo tipo de cerveja depender do sexo do consumidor, a empresa modelará suas campanhas de acordo com os diferentes mercados-alvo.

Um teste de independência trata da questão referente a se a preferência pelo tipo de cerveja (*light*, comum ou escura) independe do sexo do consumidor de cerveja (masculino ou feminino). As hipóteses para esse teste de independência são:

H_0: A preferência pelo tipo de cerveja independe do sexo do consumidor

H_a: A preferência pelo tipo de cerveja depende do sexo do consumidor

A Tabela 11.3 pode ser usada para descrever a situação que está sendo estudada. Após a identificação da população como totalmente composta de homens e mulheres consumidores de cerveja, uma amostra pode ser selecionada e cada indivíduo é solicitado a declarar sua predileção por um dos tipos de cerveja da Alber's. Cada indivíduo da amostra será classificado em uma das seis células da tabela. Por exemplo, um indivíduo pode ser um homem que prefere a cerveja comum (célula (1,2)), uma mulher que prefere a cerveja *light* (célula (2,1)), uma mulher que prefere uma cerveja escura (célula (2,3)) e assim por diante. Assim que tivermos relacionado todas as combinações possíveis de predileção pelos tipos de cerveja e pelo

Para testar se duas variáveis são independentes, uma amostra é selecionada e uma tabulação cruzada é utilizada para sintetizar simultaneamente os dados das duas variáveis.

sexo ou, em outras palavras, relacionado todas as contingências possíveis, a Tabela 11.3 passa a chamar-se **tabela de contingência.** O teste de independência usa um formato de tabela de contingência e, por essa razão, às vezes, é denominado *teste da tabela de contingência.*

Suponha que uma amostra aleatória simples de 150 consumidores de cerveja seja selecionada. Depois de degustar cada cerveja, os indivíduos da amostra foram solicitados a manifestar sua predileção, ou primeira escolha. A tabulação cruzada da Tabela 11.4 resume as respostas do estudo. Conforme observamos, os dados do teste de independência são coletados em termos de contagens ou freqüências correspondentes a cada célula ou categoria. Dos 50 indivíduos da amostra, 20 eram homens que preferiam a cerveja *light*; 40, a cerveja comum; 20, a cerveja escura; e assim por diante.

Os dados da Tabela 11.4 são as freqüências observadas das seis classes ou categorias. Se pudermos determinar as freqüências esperadas sob a hipótese de independência entre as preferências por cada tipo de cerveja e sexo do consumidor de cerveja, poderemos usar a distribuição quiquadrado para determinar se há uma diferença significativa entre as freqüências observadas e esperadas.

Tabela 11.4 Resultados amostrais da predileção por tipo de cerveja da parte de homens e mulheres consumidores de cerveja (freqüências observadas)

		Cerveja Preferida			
		Light	Comum	Escura	Total
Sexo	Masculino	20	40	20	80
	Feminino	30	30	10	70
	Total	50	70	30	150

Tabela 11.5 Freqüências esperadas se a predileção pela cerveja independer do sexo do consumidor

		Cerveja Preferida			
		Light	Comum	Escura	Total
Sexo	Masculino	26,67	37,33	16,00	80
	Feminino	23,33	32,67	14,00	70
	Total	50,00	70,00	30,00	150

As freqüências esperadas das células da tabela de contingência baseiam-se no seguinte fundamento lógico. Em primeiro lugar, supomos que a hipótese nula de independência entre a predileção por determinado tipo de cerveja e o sexo do consumidor de cerveja seja verdadeira. Então, observamos que, na amostra inteira dos 150 consumidores de cerveja, um total de 50 prefere a cerveja *light*; 70, a cerveja comum; e 30, a cerveja escura. Em termos de frações, concluímos que $^{50}/_{150} = {}^{1}/_{3}$ dos consumidores de cerveja preferem a cerveja *light*; $^{70}/_{150} = {}^{7}/_{15}$, a cerveja comum; e $^{30}/_{150} = {}^{1}/_{5}$, a cerveja escura. Se a hipótese de *independência* for válida, argumentamos que essas frações devem ser aplicáveis tanto a consumidores de cerveja homens como a consumidores de cerveja mulheres. Desse modo, sob a hipótese de independência, esperaríamos que a amostra de 80 consumidores de cerveja homens demonstre que $(^{1}/_{3})80 = 26{,}67$ preferem a cerveja *light*; $(^{7}/_{15})80 = 37{,}33$, a cerveja comum; e $(^{1}/_{5})80 = 16$, a cerveja escura. A aplicação das mesmas frações aos 70 consumidores de cerveja mulheres produz as freqüências esperadas apresentadas na Tabela 11.5.

Admitamos que e_{ij} denote a freqüência esperada correspondente à categoria da tabela de contingência na linha i, coluna j. Com essa notação, reconsideremos o cálculo da freqüência esperada para os homens (linha $i = 1$) que preferem a cerveja comum (coluna $j = 2$); ou seja, a freqüência esperada e_{12}. Seguindo o argumento anterior referente ao cálculo das freqüências esperadas, podemos mostrar que

$$e_{12} = (^{7}/_{15})80 = 37{,}33$$

Essa expressão pode ser escrita de uma maneira ligeiramente diferente:

$$e_{12} = (^{7}/_{15})80 = (^{70}/_{150})80 = \frac{(80)(70)}{150} = 37{,}33$$

Note que 80 na expressão é o número total de homens (total da linha 1), 70 é o número total de indivíduos que preferem a cerveja comum (total da coluna 2) e 150 é o tamanho total da amostra. Portanto, dizemos que

$$e_{12} = \frac{\text{(Total da Linha 1)(Total da Coluna 2)}}{\text{Tamanho da Amostra}}$$

A generalização da expressão mostra que a fórmula seguinte produz as freqüências esperadas para uma tabela de contingência no teste de independência.

FREQÜÊNCIAS ESPERADAS PARA TABELAS DE CONTINGÊNCIA SOB A HIPÓTESE DE INDEPENDÊNCIA

$$e_{ij} = \frac{\text{(Total da Linha } i\text{)(Total da Coluna } j\text{)}}{\text{Tamanho da Amostra}} \qquad \textbf{(11.9)}$$

Usando a fórmula para os consumidores de cerveja homens que preferem a cerveja escura, encontramos uma freqüência esperada igual a $e_{13} = (80)(30)/150 = 16{,}00$, como é mostrado na Tabela 11.5. Use a Equação 11.9 para verificar as outras freqüências esperadas apresentadas na Tabela 11.5.

O procedimento de teste para comparar as freqüências observadas da Tabela 11.4 com as freqüências esperadas da Tabela 11.5 é similar aos cálculos de eficiência de ajuste feitos na Seção 11.2. Especificamente, o valor χ^2 baseado nas freqüências observadas esperadas é o seguinte:

ESTATÍSTICA DE TESTE DE INDEPENDÊNCIA

$$\chi^2 = \sum_i \sum_j \frac{(f_{ij} - e_{ij})^2}{e_{ij}} \qquad \textbf{(11.10)}$$

em que

f_{ij} = freqüência observada para a categoria da tabela de contingência na linha i, coluna j.
e_{ij} = freqüência esperada para a categoria da tabela de contingência na linha i, coluna j baseada na hipótese de independência.

Observação: Com n linhas e m colunas na tabela de contingência, a estatística de teste tem uma distribuição quiquadrado com $(n - 1)(m - 1)$ graus de liberdade, desde que as freqüências esperadas sejam cinco ou mais para todas as categorias.

O somatório duplo na Equação 11.10 é usado para indicar que o cálculo deve ser feito para todas as células da tabela de contingência.

Ao revisarmos as freqüências esperadas na Tabela 11.5, vemos que as freqüências esperadas são cinco ou mais para cada categoria. Por conseguinte, prosseguimos com os cálculos da estatística de teste quiquadrado. Os cálculos necessários para calcular a estatística de teste quiquadrado para determinar se a predileção por determinado tipo de cerveja independe do gênero do consumidor são apresentados na Tabela 11.6. Notamos que o valor da estatística de teste é $\chi^2 = 6{,}12$.

O número de graus de liberdade para a distribuição quiquadrado apropriada é calculado multiplicando-se o número de linhas menos 1 pelo número de colunas menos 1. Com duas linhas e três colunas, temos $(2 - 1)(3 - 1) = 2$ graus de liberdade. À semelhança do que ocorre com o teste de eficiência de ajuste, o teste de independência rejeita H_0 se as diferenças entre as freqüências observadas e esperadas produzirem um valor grande para a estatística de teste. Assim, o teste de independência também é um teste da cauda superior. Usando a tabela quiquadrado (Tabela 3 do Apêndice B), concluímos que a área da cauda superior, ou valor p, em $\chi^2 = 6{,}12$ está entre 0,025 e 0,05. O Excel exibe o valor $p = 0{,}0468$. No nível de significância 0,05, o valor $p \leq \alpha = 0{,}05$. Rejeitamos a hipótese nula de independência e concluímos que a preferência por determinado tipo de cerveja não independe do sexo do consumidor de cerveja.

Tabela 11.6 Cálculo da estatística de teste quiquadrado para determinar se a preferência por determinado tipo de cerveja independe do sexo do consumidor

Sexo	Cerveja Preferida	Freqüência Observada (f_{ij})	Freqüência Esperada (e_{ij})	Diferença ($f_{ij} - e_{ij}$)	Quadrado da Diferença $(f_{ij} - e_{ij})^2$	Quadrado da Difference Dividido pela Freqüência Esperada $(f_{ij} - e_{ij})^2/e_{ij}$
Masculino	*Light*	20	26,67	26,67	44,44	1,67
Masculino	Comum	40	37,33	2,67	7,11	0,19
Masculino	Escura	20	16,00	4,00	16,00	1,00
Feminino	*Light*	30	23,33	6,67	44,44	1,90
Feminino	Comum	30	32,67	22,67	7,11	0,22
Feminino	Escura	10	14,00	24,00	16,00	1,14
	Total	150				$\chi^2 = 6,12$

Figura 11.1 Saída de dados (*output*) do Minitab para o teste de independência da Alber's Brewery

```
Expected counts are printed below observed counts

        Light    Regular    Dark    Total
   1      20        40        20      80
       26.67     37.33     16.00

   2      30        30        10      70
       23.33     32.67     14.00

Total     50        70        30     150

DF = 2, P-Value = 0.047
```

O teste de independência sempre é um teste unicaudal, com a região de rejeição na cauda superior da distribuição quiquadrado.

Softwares, como o Minitab e o Excel, podem simplificar o cálculo do teste de independência e fornecer o valor p para o teste. As etapas usadas para se obter os resultados de computador para um teste de independência são apresentadas nos Apêndices 11.2 e 11.3. A saída de dados do Minitab para o teste da independência da Alber's Brewery é apresentada na Figura 11.1.

Embora os resultados do teste não nos permitam tirar conclusões adicionais, podemos comparar informalmente as freqüências observadas e esperadas para obtermos uma idéia a respeito da dependência entre a predileção por um tipo de cerveja e o gênero dos consumidores. Consultemos as Tabelas 11.4 e 11.5. Vemos que os consumidores de cerveja homens têm freqüências observadas maiores que as freqüências esperadas tanto para a cerveja comum como para a cerveja escura, ao passo que os consumidores de cerveja mulheres têm uma freqüência esperada maior que a freqüência esperada somente no que se refere à cerveja *light*. Essas observações nos dão *insight* sobre as diferenças quanto à predileção por um tipo de cerveja entre os consumidores homens e mulheres.

Vamos resumir as etapas do teste de independência em uma tabela de contingência.

TESTE DE INDEPENDÊNCIA: RESUMO

1. Estabeleça as hipóteses nula e alternativa.

   H_0: A variável coluna é independente da variável linha
   H_a: A variável coluna não é independente da variável linha

2. Selecione uma amostra aleatória e registre as freqüências observadas para cada célula da tabela de contingência.
3. Use a Equação 11.9 para calcular a freqüência esperada de cada célula.
4. Use a Equação 11.10 para calcular o valor da estatística de teste.
5. Regra de rejeição:

Critério do valor p: Rejeitar H_0: se o valor $p \leq \alpha$
Critério do valor crítico: Rejeitar H_0 se $\chi^2 \geq \chi^2_\alpha$

em que α é o nível de significância, com n linhas e m colunas produzem $(n-1)(m-1)$ graus de liberdade.

NOTAS E COMENTÁRIOS

A estatística de teste para os testes quiquadrado apresentados neste capítulo requer uma freqüência esperada igual a cinco para cada categoria. Quando uma categoria tem um número menor que 5, freqüentemente é apropriado combinar duas categorias adjacentes de cinco ou mais unidades em cada uma.

Exercícios

Métodos

19. A tabela de contingência 2 × 3 apresentada a seguir contém as freqüências observadas correspondentes a uma amostra de tamanho 200. Teste a independência das variáveis linha e coluna usando o teste χ^2 com $\alpha = 0{,}05$.

	Variável Coluna		
Variável Linha	A	B	C
P	20	44	50
Q	30	26	30

AUTOTESTE

20. A tabela de contingência 3 × 3 apresentada a seguir contém as freqüências observadas correspondentes a uma amostra de tamanho 240. Teste a independência das variáveis linha e coluna usando χ^2 com $\alpha = 0{,}05$.

	Variável Coluna		
Variável Linha	A	B	C
P	20	30	20
Q	30	60	25
R	10	15	30

Aplicações

21. Uma das questões do Estudo dos Assinantes (Subscriber Study, da *Business Week*) foi: "Nos últimos doze meses, ao fazer viagens de negócios, qual tipo de passagem aérea você comprou mais freqüentemente?". Os dados obtidos são apresentados na seguinte tabela de contingência:

	Tipo de Vôo	
Tipo de Passagem	Vôos Domésticos	Vôos Internacionais
Primeira-classe	29	22
Classe *Business*/Executiva	95	121
Full fare economy[1]/Classe turística	518	135

AUTOTESTE

Use $\alpha = 0{,}05$ e teste a independência do tipo de vôo e do tipo de passagem. Qual é a sua conclusão?

22. Em um estudo sobre a fidelidade à marca na indústria automobilística, compradores de carros novos foram solicitados a responder se a marca de seu carro era a mesma do carro que possuíam anteriormente (*Business Week*, 8 de maio de 2000). O detalhamento das 600 respostas mostra a fidelidade à marca relativa a carros nacionais, europeus e asiáticos.

[1] NT: *Full fare economy* – Nas viagens aéreas, as classes *full fare economy* e turística são, ambas, econômicas. A diferença entre uma e outra é que a classe *full fare economy* é um pouco mais cara e geralmente dá certos direitos, como reembolsos, transferências, troca de dias e horários etc.

Comprou	Fabricante Nacional	Europeu	Asiático
A Mesma Marca	125	55	68
Uma Marca Diferente	140	105	107

a. Teste uma hipótese para determinar se a fidelidade à marca independe do fabricante. Use $\alpha = 0{,}05$. Qual é a sua conclusão?
b. Se uma diferença significativa for encontrada, qual fabricante parece contar com a maior fidelidade à marca?

23. Em virtude dos aumentos percentuais anuais de dois dígitos no custo dos seguros-saúde, um número cada vez maior de trabalhadores provavelmente ficará sem a cobertura de um seguro-saúde (*USA Today*, 23 de janeiro de 2004). Os dados amostrais seguintes fornecem uma comparação dos trabalhadores de pequenas, médias e grandes empresas que possuem e as que não possuem cobertura de um seguro-saúde. Para a finalidade desse estudo, denominamos pequenas empresas as que têm menos de 100 funcionários; médias empresas são as que possuem de 100 a 999 funcionários e grandes empresas são as empresas que contam com mais de mil funcionários. Dados amostrais referentes a 50 funcionários de pequenas empresas, 75 de médias empresas e 100 de grandes empresas foram registrados.

Tamanho da Empresa	Seguro-Saúde Sim	Não	Total
Pequena	36	14	50
Média	65	10	75
Grande	88	12	100

a. Realize um teste de independência para determinar se a cobertura de seguro-saúde dos funcionários independe do tamanho da empresa. Use $\alpha = 0{,}05$. Qual é o valor p e qual é a sua conclusão?
b. O artigo do *USA Today* indicou que os funcionários de pequenas empresas têm mais probabilidade de ficar sem cobertura de um seguro-saúde. Use porcentagens baseadas nos dados acima para dar suporte a essa conclusão.

24. Um estudo realizado pelo Public Interest Research Group (PIRG) do Estado de Washington revelou que 46% dos estudantes universitários que fazem cursos de tempo integral realizam seus trabalhos acadêmicos durante mais de 25 horas por semana. O estudo do PIRG apresentou dados sobre os efeitos dos trabalhos acadêmico sobre a obtenção do diploma (*USA Today*, 17 de abril de 2002). Uma amostra de 200 estudantes incluiu 90 que realizavam trabalhos acadêmicos de 1 a 15 horas por semana; 60, de 16 a 24 horas por semana; e 50, de 25 a 34 horas por semana. O número de estudantes da amostra que indicaram que seus trabalhos tinham um efeito positivo, nenhum efeito ou um efeito negativo sobre a obtenção de seus diplomas são os seguintes:

Horas de Trabalho por Semana	Efeito sobre a Obtenção do Diploma Positivo	Nenhum	Negativo	Total
1 a 15 horas	26	50	14	90
16 a 24 horas	16	27	17	60
25 a 34 horas	11	19	20	50

a. Realize um teste de independência para determinar se o efeito sobre a obtenção do diploma independe das horas de trabalho por semana. Use $\alpha = 0{,}05$. Qual é o valor p e qual é a sua conclusão?
b. Use porcentagens de linha para conhecer melhor como os trabalhos acadêmicos afetam a obtenção do diploma. Qual é a sua conclusão?

25. O apelo negativo é reconhecido como um método eficaz de persuasão na propaganda. Um estudo publicado em *The Journal of Advertising* relatou os resultados de uma análise de conteúdo de publicidades com apelos à culpa e ao medo veiculados em 24 revistas. O número de anúncios com apelos à culpa e ao medo que apareceram em tipos de revista selecionados refere-se ao:

	Tipo de Apelo	
Tipo de Revista	Número de Anúncios com Apelos à Culpa	Número de Anúncios com Apelos ao Medo
Notícias e opinião	20	10
Editoria geral	15	11
Orientada à família	30	19
Negócios/finanças	22	17
Orientada à mulher	16	14
Afro-americana	12	15

Use o teste de independência quiquadrado com o nível de significância 0,01 para analisar os dados. Qual é a sua conclusão?

26. O comércio faz colocação de pedidos on-line em um número cada vez maior. O Performance Measurement Group coletou dados sobre os custos das encomendas eletrônicas atendidas corretamente pela indústria (*Investor's Business Daily*, 8 de maio de 2000). Suponha que uma amostra de 700 encomendas eletrônicas tenha produzido os seguintes resultados:

	Indústria			
Pedido	Produtos Farmacêuticos	Itens de Consumo	Computadores	Equipamentos de Telecomunicações
Corretos	207	136	151	178
Incorretos	3	4	9	12

a. Teste uma hipótese para determinar se a exatidão em termos de atendimento do pedido independe da indústria. Use $\alpha = 0,05$. Qual é a sua conclusão?
b. Qual indústria tem a porcentagem mais elevada de exatidão no atendimento dos pedidos?

27. A National Sleep Foundation utilizou uma pesquisa para determinar se as horas de sono por noite independem da idade (*Newsweek*, 19 de janeiro de 2004). Os dados a seguir apresentam as horas de sono por noite de uma amostra de indivíduos com menos de 49 anos e de uma amostra de indivíduos com mais de 50 anos.

	Horas de Sono				
Idade	Menos de 6	de 6 a 6,9	de 7 a 7,9	8 ou mais	Total
Menos de 49 anos	38	60	77	65	240
Mais de 50 anos	36	57	75	92	260

a. Realize um teste de independência para determinar se as horas de sono por noite independem da idade. Use $\alpha = 0,05$. Qual é o valor p e qual é a sua conclusão?
b. Qual é a sua estimativa da porcentagem de pessoas que dormem menos de 6 horas, de 6 a 6,9 horas, de 7 a 7,9 horas e de 8 ou mais horas por noite?

Resumo

Neste capítulo, descrevemos procedimentos estatísticos que envolvem proporções e o teste de independência da tabela de contingência de duas variáveis. Na primeira seção, comparamos uma proporção de uma população com a mesma proporção de outra. Descrevemos como construir uma estimação por intervalo da diferença entre as proporções e como realizar um teste de hipóteses para saber se a diferença entre as proporções era estatisticamente significativa.

Na segunda seção, concentramo-nos em uma única população multinomial. Ali, vimos como realizar testes de hipóteses para determinar se as proporções amostrais correspondentes às categorias da população multinomial eram significativamente diferentes dos valores tomados como hipótese. O teste quiquadrado de eficiência de ajuste foi utilizado para fazer a comparação. A seção final foi preenchida com testes de independência para duas variáveis. Um teste de independência para duas variáveis é uma extensão da metodologia empregada no teste de eficiência de ajuste para uma população multinomial. Uma tabela de contingência é usada para determinar as freqüências observadas e esperadas. Então, é calculado um valor quiquadrado. Valores quiquadrado grandes, causados por diferenças grandes entre as freqüências observadas e esperadas, levam à rejeição da hipótese nula de independência.

Glossário

Estimador agrupado de p Um estimador de uma proporção populacional obtido calculando-se a média ponderada das proporções amostrais obtidas de duas amostras independentes.

População multinomial Uma população na qual cada elemento é designado a uma e somente uma de diversas categorias. A distribuição multinomial amplia a distribuição binomial de dois para três ou mais resultados.

Teste de eficiência de ajuste Um teste estatístico realizado para determinar se convém rejeitar uma distribuição de probabilidade hipotética referente a uma população.

Tabela de contingência Uma tabela usada para resumir as freqüências observadas e esperadas de um teste de independência.

Fórmulas-Chave

Estimador por Ponto da Diferença Entre Duas Proporções Populacionais

$$\bar{p}_1 - \bar{p}_2 \tag{11.1}$$

Erro Padrão de $\bar{p}_1 - \bar{p}_2$

$$\sigma_{\bar{p}_1-\bar{p}_2} = \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}} \tag{11.2}$$

Estimação por Intervalo da Diferença Entre Duas Proporções Populacionais

$$\bar{p}_1 - \bar{p}_2 \pm z_{\alpha/2}\sqrt{\frac{\bar{p}_1(1-\bar{p}_1)}{n_1} + \frac{\bar{p}_2(1-\bar{p}_2)}{n_2}} \tag{11.4}$$

Erro Padrão de $\bar{p}_1 - \bar{p}_2$ Quando $p_1 = p_2 = p$

$$\sigma_{\bar{p}_1-\bar{p}_2} = \sqrt{p(1-p)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)} \tag{11.5}$$

Estimador Agrupado de p Quando $p_1 = p_2 = p$

$$\bar{p} = \frac{n_1\bar{p}_1 + n_2\bar{p}_2}{n_1 + n_2} \tag{11.6}$$

Estatística de Teste para Testes de Hipóteses sobre $p_1 - p_2$

$$z = \frac{(\bar{p}_1 - \bar{p}_2)}{\sqrt{\bar{p}(1-\bar{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \tag{11.7}$$

Estatística de Teste da Eficiência de Ajuste

$$\chi^2 = \sum_{i=1}^{k} \frac{(f_i - e_i)^2}{e_i} \tag{11.8}$$

Freqüências Esperadas para Tabelas de Contingência sob a Hipótese de Independência

$$e_{ij} = \frac{(\text{Total da Linha } i)(\text{Total da Coluna } j)}{\text{Tamanho da Amostra}} \tag{11.9}$$

Estatística de Teste de Independência

$$\chi^2 = \sum_i \sum_j \frac{(f_{ij} - e_{ij})^2}{e_{ij}} \tag{11.10}$$

Exercícios Suplementares

28. A Jupiter Media utilizou uma pesquisa para determinar como as pessoas usam o tempo livre. Assistir à televisão foi a atividade mais popular escolhida tanto pelos homens quanto pelas mulheres (*The Wall*

Street Journal, 26 de janeiro de 2004). A proporção de homens e a proporção de mulheres que escolheram assistir à televisão como a atividade de lazer mais popular podem ser estimadas em decorrência dos seguintes dados amostrais.

Gênero	Tamanho da Amostra	Assistir à Televisão
Homens	800	248
Mulheres	600	156

a. Estabeleça as hipóteses que podem ser usadas para testar a diferença entre a proporção da população de homens e a proporção da população de mulheres que escolheram assistir à televisão como a atividade de lazer mais popular.

b. Qual é a proporção amostral de homens que escolheram assistir à televisão como a atividade de lazer mais popular? Qual é a proporção amostral de mulheres?

c. Realize um teste de hipóteses e calcule o valor p. Ao nível de significância 0,05, qual é a sua conclusão?

d. Qual é a margem de erro e a estimação por intervalo de confiança de 95% entre as proporções populacionais?

29. Uma grande empresa de seguros de automóvel selecionou amostras de segurados do sexo masculino solteiros e casados e registrou o número dos que fizeram uma reclamação de seguro ao longo dos três últimos anos.

Segurados Solteiros	Segurados Casados
$n_1 = 400$	$n_2 = 900$
Número dos que fizeram uma reclamação de seguro = 76	Número dos que fizeram uma reclamação de seguro = 90

a. Use $\alpha = 0{,}05$. Teste para determinar se os índices de reclamação de seguros diferem entre os segurados solteiros e casados.

b. Forneça um intervalo de confiança de 95% correspondente à diferença entre as proporções das duas populações.

30. Exames médicos foram realizados para se conhecer melhor a tuberculose resistente a medicamentos. De 142 casos examinados em New Jersey, nove foram considerados resistentes a medicamentos. Dos 268 casos examinados no Texas, cinco foram considerados resistentes a medicamentos. Esses dados sugerem uma diferença estatisticamente significativa entre a proporção de casos resistentes a medicamentos nos dois estados? Use um nível de significância 0,02. Qual é o valor p e qual é a sua conclusão?

31. Em julho de 2001, a Harris Ad Track Research Service realizou uma pesquisa para avaliar a eficácia de uma grande campanha publicitária das câmaras Kodak (*USA Today*, 27 de agosto de 2001). Em uma amostra de 430 entrevistados, 38% consideraram os anúncios muito eficazes. Em outra amostra de 285 entrevistados quanto a outras campanhas publicitárias, 23% consideraram os anúncios muito eficazes.

a. Estime o número de entrevistados que consideraram os anúncios da Kodak muito eficazes e o número de entrevistados que acharam os outros anúncios muito eficazes.

b. Forneça um intervalo de confiança de 95% para a diferença entre as proporções.

c. Com base nos resultados que obteve no item (b), você acredita que a campanha publicitária da Kodak é mais eficaz que a maioria das campanhas publicitárias?

32. Em junho de 2001, 38% dos gerentes de fundos financeiros pesquisados acreditavam que o índice do núcleo inflacionário seria mais alto em um ano. Um mês depois, uma pesquisa idêntica revelou que 22% dos gerentes de fundos financeiros esperavam que o índice do núcleo inflacionário seria mais alto em um ano (*Global Research Highlights*, Merrill Lynch, 20 de julho de 2001). Suponha que o tamanho da amostra tenha sido 200 tanto na pesquisa de junho como na de julho.

a. Desenvolva uma estimação por ponto da diferença entre as proporções de junho e julho de gerentes de fundos financeiros que achavam que o índice do núcleo inflacionário seria mais alto em um ano.

b. Desenvolva hipóteses de tal forma que a rejeição da hipótese nula nos permita concluir que as expectativas de inflação diminuíram entre junho e julho.

c. Realize um teste das hipóteses dos itens (a) e (b) usando $\alpha = 0{,}01$. Qual é a sua conclusão?

33. Sete por cento dos investidores de fundos mútuos classificam as obrigações ao portador de "muito seguras"; 58%, de "relativamente seguras"; 24%, de "não muito seguras"; 4%, de "absolutamente inseguras"; e 7% "não sabem". Uma pesquisa de opinião promovida pela *Business Week*/Harris per-

guntou a 529 investidores de fundos mútuos como eles avaliariam os títulos privados em termos de segurança. As respostas foram as seguintes:

Avaliação da Segurança	Freqüência
Muito seguros	48
Relativamente seguros	323
Não muito seguros	79
Absolutamente inseguros	16
Não sabem	63
Total	529

A postura que os investidores de fundos mútuos possuem em relação aos títulos privados é diferente daquela que têm quanto às obrigações ao portador? Sustente sua conclusão com um teste estatístico. Use $\alpha = 0{,}01$.

34. Desde 2000, o Toyota Camry, o Honda Accord e o Ford Taurus são os três veículos de passageiros mais vendidos nos Estados Unidos. Com base nos dados de vendas de 2003, as participações no mercado entre os três mais vendidos são: Toyota Camry, 37%; Honda Accord, 34%; e Ford Taurus, 29% (*The World Almanac*, 2004). Suponha que uma amostra de 1.200 vendas de carros de passageiros durante o primeiro trimestre de 2004 apresente o seguinte:

Carro de Passageiros	Unidades Vendidas
Toyota Camry	480
Honda Accord	390
Ford Taurus	330

Esses dados podem ser usados para concluirmos que as participações no mercado entre os três carros de passageiros mais vendidos se alteraram durante o primeiro trimestre de 2004? Qual é o valor p? Use o nível de significância 0,05. Qual é a sua conclusão?

35. Uma autoridade regional de trânsito está preocupada com o número de ciclistas em uma de suas rotas de ônibus. Ao planejar a rota, a suposição é de que o número de ciclistas seja o mesmo todos os dias, de segunda a sexta-feira. Usando os dados apresentados a seguir, faça um teste com $\alpha = 0{,}05$ para determinar se a suposição da autoridade de trânsito está correta.

Dia	Número de Ciclistas
Segunda-feira	13
Terça-feira	16
Quarta-feira	28
Quinta-feira	17
Sexta-feira	16

36. Os resultados da pesquisa intitulada Annual Job Satisfaction Survey, da *Computerworld*, revelaram que 28% dos gerentes de Sistemas de Informação (SI) estão muito satisfeitos com seus empregos, 46% estão relativamente satisfeitos, 12% não estão nem satisfeitos nem insatisfeitos, 10% estão relativamente insatisfeitos, e 4% estão muito insatisfeitos. Suponha que uma amostra de 500 programadores tenha fornecido os seguintes resultados.

Categoria	Número de Entrevistados
Muito satisfeitos	105
Relativamente satisfeitos	235
Nem satisfeitos nem insatisfeitos	55
Relativamente insatisfeitos	90
Muito insatisfeitos	15

Use $\alpha = 0{,}05$ e faça um teste para determinar se a satisfação com o trabalho da parte dos programadores é diferente da satisfação com o trabalho da parte dos gerentes de sistemas de informação.

37. Uma amostra de peças forneceu os dados da seguinte tabela de contingência sobre a qualidade das peças por turno de produção.

Turno	Número de Peças com Boa Qualidade	Número de Peças Defeituosas
Primeiro	368	32
Segundo	285	15
Terceiro	176	24

Use $\alpha = 0{,}05$ e teste a hipótese de que a qualidade das peças independe do turno de produção. Qual é a sua conclusão?

38. Um Estudo dos Assinantes (Subscriber Study) do *The Wall Street Journal* publicou dados sobre a situação de emprego dos assinantes. Os resultados da amostra correspondentes aos assinantes das edições da região leste e oeste do país aparecem na tabela a seguir:

	Região	
Situação de Emprego	**Edição do Leste**	**Edição do Oeste**
Tempo integral	1.105	574
Tempo parcial	31	15
Autônomo/consultor	229	186
Não empregado	485	344

Use $\alpha = 0{,}05$ e teste a hipótese de que a situação de emprego independe da região do país. Qual é a sua conclusão?

39. Uma instituição de empréstimo forneceu os seguintes dados sobre aprovações de crédito em quatro departamentos de crédito. Use $\alpha = 0{,}05$ e faça um teste para determinar se a decisão de aprovar o crédito independe do diretor de crédito que analisa o pedido de empréstimo:

	Decisão de Aprovação do Empréstimo	
Diretor de Crédito	**Aprovado**	**Recusado**
Miller	24	16
McMahon	17	13
Games	35	15
Runk	11	9

40. Dados sobre o estado civil de homens e mulheres com idades de 20 a 29 anos foram obtidos como parte de uma pesquisa nacional. Os resultados de uma amostra de 350 homens e 400 mulheres são os seguintes:

		Estado Civil	
Gênero	**Solteiro**	**Casado**	**Divorciado**
Homem	234	106	10
Mulheres	216	168	16

a. Use $\alpha = 0{,}01$ e teste a independência entre o estado civil e o gênero. Qual é a sua conclusão?

b. Resuma a porcentagem de cada categoria de estado civil relativa aos homens e às mulheres.

41. O Barna Research Group coletou dados que mostram a freqüência à igreja por faixa etária (*USA Today*, 20 de novembro de 2003). Use os dados amostrais para determinar se a freqüência à igreja independe da idade. Use um nível de significância 0,05. Qual é a sua conclusão? Qual conclusão você pode tirar a respeito da freqüência à igreja à medida que as pessoas se tornam mais velhas?

	Freqüência à Igreja		
Idade	**Sim**	**Não**	**Total**
20 a 29	31	69	100
30 a 39	63	87	150
40 a 49	94	106	200
50 a 59	72	78	150

42. Um vendedor realiza quatro contatos de vendas por dia. Uma amostra de 100 dias fornece as seguintes freqüências de volumes de vendas.

Número de Vendas	Freqüência Observada (em dias)
0	30
1	32
2	25
3	10
4	3
Total	100

Registros mostram que são realizadas vendas em 30% de todos os contatos de vendas. Supondo contatos de vendas independentes, o número de vendas por dia deve seguir uma distribuição binomial. A função binomial de probabilidade apresentada no Capítulo 5 é

$$f(x) = \frac{n!}{x!(n-x)!} p^x(1-p)^{n-x}$$

Para esse exercício, suponha que a população tenha uma distribuição binomial com $n = 4$, $p = 0{,}30$ e $x = 0, 1, 2, 3$ e 4.

a. Calcule as freqüências esperadas para $x = 1, 2, 3$ e 4 usando a função binomial de probabilidade. Se necessário, combine categorias para satisfazer o requisito de a freqüência esperada ser cinco ou mais de todas as categorias.
b. Use um teste de eficiência de ajuste para determinar se a hipótese de uma distribuição binomial deve ser rejeitada. Use $\alpha = 0{,}05$. Uma vez que nenhum parâmetro da distribuição binomial foi estimado a partir dos dados amostrais, os graus de liberdade são $k - 1$ quando k é o número de categorias.

Estudo de Caso – Programa Bipartidário de Reforma

Em um estudo realizado pela Zogby International para o periódico *Democrat and Chronicle*, mais de 700 nova-iorquinos foram consultados para determinar se o governo do estado de Nova York é competente. Os entrevistados integrantes da pesquisa foram solicitados a responder perguntas que envolviam reduções salariais aos deputados estaduais, restrições aos lobistas, limitações de tempo de mandato para os parlamentares e possibilidade de os cidadãos de cada estado sugerirem temas diretamente para serem votados. Os resultados referentes a diversas reformas propostas obtiveram amplo apoio entre todos os níveis demográficos e políticos.

Suponha que uma pesquisa de acompanhamento de 100 indivíduos que vivem na região oeste de Nova York tenha sido realizada. A filiação partidária (Democrata, Republicano, Independente) de cada indivíduo foi registrada, bem como as respostas individuais às três perguntas seguintes:

1. A remuneração dos parlamentares deve ser reduzida em correspondência a cada dia de atraso do orçamento estadual?

   Sim _____ Não_____

2. Deve haver mais restrições aos lobistas?

   Sim _____ Não _____

3. Deve haver limitação do tempo de mandato exigindo que os parlamentares permaneçam um número fixo de anos no cargo?

   Sim _____ Não _____

ARQUIVO DA INTERNET NYReform

As respostas foram codificadas usando 1 para as respostas *sim* e 2 para as respostas *não*. O conjunto de dados (*data set*) completo está disponível no arquivo intitulado NYReform no site.

Relatório Administrativo

1. Use a estatística descritiva para resumir os dados desse estudo. Quais são suas conclusões preliminares sobre a independência da resposta (Sim ou Não) e filiação partidária em relação a cada uma das três perguntas da pesquisa?

2. Com respeito à questão 1, teste a independência da resposta (Sim ou Não) e a filiação partidária. Use $\alpha = 0{,}05$.
3. Com relação à questão 2, teste a independência da resposta (Sim ou Não) e a filiação partidária. Use $\alpha = 0{,}05$.
4. Referente à questão 3, teste a independência da resposta (Sim ou Não) e a filiação partidária. Use $\alpha = 0{,}05$.
5. Parece haver amplo apoio à reforma entre todas as tendências políticas? Explique.

Apêndice 11.1 – Inferências sobre Duas Proporções Populacionais com o Minitab

ARQUIVO DA INTERNET
TaxPrep

Descrevemos o uso do Minitab para desenvolver estimações por intervalo e realizar testes de hipóteses sobre as diferenças entre duas proporções. Usaremos os dados de erros de preenchimento do imposto de renda apresentados na Seção 11.1. Os resultados amostrais referentes a 250 declarações do imposto de renda preparados no escritório 1 estão na coluna C1 e os resultados amostrais de 300 declarações preparadas no escritório 2 estão na coluna C2. Sim denota que um erro foi detectado na declaração do imposto de renda e Não indica que nenhum erro foi encontrado. O procedimento que descrevemos fornece tanto uma estimação por intervalo de 90% de confiança da diferença entre as duas proporções populacionais como os resultados do teste de hipóteses para H_0: $p_1 - p_2 = 0$ contra H_a: $p_1 - p_2 \neq 0$.

Etapa 1. Selecione o menu **Stat**
Etapa 2. Escolha a opção **Basic Statistics**
Etapa 3. Escolha a opção **2 Proportions**
Etapa 4. Quando a caixa de diálogo 2 Proportions (Test and Confidence Interval) aparecer:
 Selecione **Samples in different columns**
 Digite C1 na caixa **First**
 Digite C2 na caixa **Second**
 Selecione **Options**
Etapa 5. Quando a caixa de diálogo 2 Proportions-Options aparecer:
 Digite 90 na caixa **Confidence level**
 Digite 0 na caixa **Test difference**
 Digite not equal (não igual) na caixa **Alternative**
 Selecione **Use pooled estimate of p for test**
 Dê um clique em **OK**
Etapa 6. Dê um clique em OK

A etapa 5 pode ser modificada para produzir diferentes níveis de valores hipotéticos e diferentes formas das hipóteses.

No exemplo de preenchimento do imposto de renda, os dados são qualitativos. Sim e Não são usados para indicar se há algum erro. Nos módulos que envolvem proporções, o Minitab calcula proporções para a resposta que aparece em segundo lugar na ordem alfabética. Desse modo, no exemplo de preenchimento do imposto de renda, o Minitab calcula a proporção de respostas Sim, a qual é a proporção que queríamos.

Se a classificação em ordem alfabética do Minitab não calcular a proporção da resposta que nos interessa, podemos resolver isso. Selecionamos qualquer célula na coluna de dados, vamos à barra de menus do Minitab e selecionamos Editor > Column > Value Order. Essa seqüência oferecerá a opção de introduzirmos uma ordem de classificação especificada pelo usuário. Simplesmente, certifique-se de que a resposta de interesse esteja relacionada na caixa *define-an-order*. A rotina 2 Proportion do Minitab fornecerá então o intervalo de confiança e os resultados do teste de hipóteses referentes à proporção populacional de interesse.

Finalmente, notamos que a rotina 2 Proportion do Minitab usa um procedimento computacional diferente do procedimento descrito no texto. Assim, pode-se esperar que a saída de dados (*output*) do Minitab forneça estimações por intervalo e valores *p* ligeiramente diferentes. Entretanto, os resultados dos dois métodos devem estar próximos entre si, e espera-se que forneçam a mesma interpretação e conclusão.

Apêndice 11.2 – Testes de Eficiência de Ajuste e de Independência com o Minitab

Teste de Eficiência de Ajuste

Esse procedimento do Minitab pode ser usado para um teste de eficiência de ajuste de uma distribuição multinomial. O usuário precisa obter as freqüências observadas, calcular as freqüências esperadas e inserir tanto as freqüências observadas como as freqüências esperadas em uma planilha do Minitab. A Coluna C1 é rotulada como Observed e contém as freqüências observadas. A Coluna C2 é rotulada como Expected e contém as freqüências esperadas. Use o exemplo da Scott Marketing Research apresentado na Seção 11.2, abra uma planilha do Minitab, digite as freqüências observadas 48, 98 e 54 na coluna C1 e digite as freqüências esperadas 60, 100 e 40 na coluna C2. As etapas do Minitab para o teste de eficiência de ajuste são as seguintes:

Etapa 1. Selecione o menu **Calc**
Etapa 2. Escolha a opção **Calculator**
Etapa 3. Quando a caixa de diálogo Calculator aparecer:
  Digite ChiSquare na caixa **Store result in variable**
  Digite Sum((C1-C2)**2/C2) na caixa **Expression**
  Dê um clique em **OK**
Etapa 4. Selecione o menu **Calc**
Etapa 5. Escolha **Probability Distributions**
Etapa 6. Escolha **Chi-Square**
Etapa 7. Quando a caixa de diálogo Chi-Square Distribution aparecer:
  Selecione **Cumulative probability**
  Digite 2 na caixa **Degrees of freedom**
  Selecione **Input column** e digite ChiSquare na caixa
  Dê um clique em **OK**

A saída de dados (*output*) do Minitab fornece a probabilidade cumulativa 0,9745, a qual está na área sob a curva à esquerda de $\chi^2 = 7{,}34$. A área restante na cauda superior é o valor p. Desse modo, obtemos valor $p = 1 - 0{,}9745 = 0{,}0255$.

Teste de Independência

Iniciamos com uma nova planilha do Minitab e inserimos os dados de freqüência observada correspondentes ao exemplo da Alber's Brewery da Seção 11.3 nas colunas 1, 2 e 3, respectivamente. Dessa forma, inserimos as freqüências observadas correspondentes à preferência pela cerveja *light* (20 e 30) na coluna C1, as freqüências observadas correspondentes à preferência pela cerveja comum (40 e 30) em C2 e as freqüências observadas correspondentes à preferência pela cerveja escura (20 e 10) em C3. As etapas do Minitab para o teste de independência são as seguintes:

Etapa 1. Selecione o menu **Stat**
Etapa 2. Selecione **Tables**
Etapa 3. Escolha a opção **Chi-Square Test (Table in Worksheet)**
Etapa 4. Quando a caixa de diálogo Chi-Square Test aparecer:
  Digite C1-C3 na caixa **Columns containing the table**
  Dê um clique em **OK**

Apêndice 11.3 – Testes de Eficiência de Ajuste e de Independência com o Excel*

Teste de Eficiência de Ajuste

ARQUIVO DA INTERNET
FitTest

Esse procedimento do Excel usa um teste de eficiência de ajuste para uma distribuição multinomial. O usuário precisa obter as freqüências observadas, calcular as freqüências esperadas e inserir tanto as freqüências observadas como as freqüências esperadas em uma planilha do Excel.

* Não há rotinas disponíveis para se fazer inferências sobre a diferença entre duas proporções populacionais.

As freqüências observadas e esperadas correspondentes ao exemplo da Scott Marketing Research da Seção 11.2 são inseridas nas colunas A e B, como mostra a Figura 11.2. A estatística de teste $\chi^2 = 7{,}34$ é calculada na coluna D. Com $k = 3$ categorias, o usuário insere os graus de liberdade $k - 1 = 3 - 1 = 2$ na célula D11. A função DIST.QUI fornece o valor p na célula D13. A planilha em segundo plano exibe as fórmulas contidas nas células.

Teste de Independência

O procedimento do Excel para o teste de independência exige que o usuário obtenha as freqüências observadas e as insira na planilha. O exemplo da Alber's Brewery da Seção 11.3 fornece as freqüências observadas, as quais são inseridas nas células B7 a D8, como é mostrado na planilha da Figura 11.3. As fórmulas contidas em células apresentadas na planilha em segundo plano mostram o procedimento usado para calcular as freqüências esperadas. Com duas linhas e três colunas, o usuário insere os graus de liberdade $(2 - 1)(3 - 1) = 2$ na célula E22. A função TESTE.QUI fornece o valor p na célula E24.

Figura 11.2 Planilha do Excel para o teste de eficiência de ajuste da Scott Marketing Research

	A	B	C	D	E
1	Teste de eficiência de ajuste				
2					
3	Freqüência	Freqüência			
4	Observada	Esperada		Cálculos	
5	48	60		=(A5-B5)^2/B5	
6	98	100		=(A6-B6)^2/B6	
7	54	40		=(A7-B7)^2/B7	
8					
9			Estatística de Teste	=SOMA(D5:D7)	
10					
11			Graus de Liberdade	2	
12					
13			Valor de p	=DIST.QUI(D9,D11)	
14					

	A	B	C	D	E
1	Teste de eficiência de ajuste				
2					
3	Freqüência	Freqüência			
4	Observada	Esperada		Cálculos	
5	48	60		2,40	
6	98	100		0,04	
7	54	40		4,90	
8					
9			Estatística de Teste	7,34	
10					
11			Graus de Liberdade	2	
12					
13			Valor de p	0,0255	
14					

Figura 11.3 Planilha do Excel para o teste de independência da Alber's Brewery

	A	B	C	D	E	F
1	**Teste de Independência**					
2						
3	**Freqüência Observada**					
4						
5			**Cerveja Preferida**			
6	**Sexo**	*Light*	Comum	Escura	Total	
7	Masculino	20	40	20	=SOMA(B7:D7)	
8	Feminino	30	30	10	=SOMA(B8:D8)	
9	Total	=SOMA(B7:B8)	=SOMA(C7:C8)	=SOMA(D7:D8)	=SOMA(E7:E8)	
10						
11						
12	**Freqüência Esperada**					
13						
14			**Cerveja Preferida**			
15	**Sexo**	Light	Comum	Escura	Total	
16	Masculino	=E7*B$9/$E$9	=E7*C$9/$E$9	=E7*D$9/$E$9	=SUM(B16:D16)	
17	Feminino	=E8*B$9/$E$9	=E8*C$9/$E$9	=E8*D$9/$E$9	=SUM(B17:D17)	
18	Total	=SUM(B16:B17)	=SUM(C16:C17)	=SUM(D16:D17)	=SUM(E16:E17)	
19						
20				**Estatística de Teste**	=INV.QUI(E24,E22)	
21						
22				**Graus de Liberdade**	2	
23						
24				**Valor de *p***	=TESTE.QUI(C7:E8,C16:E17	
25						

	A	B	C	D	E	F
1	**Teste de Independência**					
2						
3	**Freqüência Observada**					
4						
5			**Cerveja Preferida**			
6	**Sexo**	*Light*	Comum	Escura	Total	
7	Masculino	20	40	20	80	
8	Feminino	30	30	10	70	
9	Total	50	70	30	150	
10						
11						
12	**Freqüência Esperada**					
13						
14			**Cerveja Preferida**			
15	**Sexo**	*Light*	Regular	Escura	Total	
16	Masculino	26,67	37,33	16	80	
17	Feminino	23,33	32,67	14	70	
18	Total	50	70	30	150	
19						
20				**Estatística de Teste**	6,12	
21						
22				**Graus de Liberdade**	2	
23						
24				**Valor de *p***	0,0468	
25						

Regressão Linear Simples

ESTATÍSTICA NA PRÁTICA

ALLIANCE DATA SYSTEMS*
Dallas, Texas

A Alliance Data Systems (ADS) oferece processamento de transações, serviços de crédito e serviços de marketing a clientes da crescente indústria do gerenciamento das relações com o cliente – *customer relationship management* (CRM). Os clientes da ADS concentram-se em quatro indústrias: varejo, petróleo/lojas de conveniência, serviços públicos e transportes. Em 1983, a Alliance começou a oferecer serviços de processamento de crédito end-to-end[1] para indústrias varejistas, petrolíferas e de restaurantes; atualmente, a empresa emprega mais de 6.500 funcionários que prestam serviços a clientes de todas as partes do mundo. Operando mais de 140 mil terminais de ponto-de-venda somente nos Estados Unidos, a ADS processa mais de 2,5 bilhões de transações anualmente. A empresa classifica-se em segundo lugar nos Estados Unidos no setor de prestação de serviços de crédito private label,[2] representando 49 programas private label que contam com aproximadamente 72 milhões de portadores de cartões de crédito. Em 2001, a ADS fez uma oferta pública inicial e agora está relacionada na Bolsa de Valores de Nova York.

Como um dos seus serviços de marketing, a ADS projeta campanhas e promoções por mala direta. Considerando que seu banco de dados contém informações sobre os hábitos de compra de mais de 100 milhões de consumidores, a ADS pode visar aos consumidores que mais provavelmente se beneficiarão de uma promoção por mala direta. O Analytical Development Group utiliza análise de regressão para construir

* Os autores agradecem a Philip Clemance, diretor de desenvolvimento da Alliance Data Systems, por fornecer esta "Estatística na Prática".

[1] NT: *End-to-end* – Proteção por criptografia de uma informação veiculada por meio de um sistema de telecomunicações do ponto de origem até o ponto de destino.

[2] NT: *Private label* – Os *private labels* são operações de financiamento destinadas a pessoas físicas realizadas através de cartão de crédito emitido por empresa do ramo de comércio ou serviços.

modelos que possam medir e prever a receptividade dos clientes a campanhas de marketing direto. Alguns modelos de regressão prevêem a probabilidade de os indivíduos que recebem uma promoção efetuarem uma compra, e outros prevêem a quantia gasta por esses consumidores ao efetuarem uma compra.

Em determinada campanha, uma loja de varejo queria atrair novos clientes. Para prever o efeito da campanha, os analistas da ADS selecionaram uma amostra do banco de dados de consumidores, enviaram matérias promocionais a indivíduos selecionados e depois coletaram dados sobre as transações indicadas nas respostas dadas pelo consumidores. Foram coletados dados amostrais sobre o valor da compra efetuada pelos clientes que responderam à campanha, além de uma série de variáveis específicas ao cliente que eram consideradas úteis na previsão das vendas. A variável específica ao cliente que mais contribuiu para prever a quantia comprada foi o valor total das compras de crédito nas lojas relacionadas durante os últimos 39 meses.

Os analistas da ADS desenvolveram uma equação de regressão estimada que relaciona o valor da compra com a quantia gasta nas lojas relacionadas:

$$\hat{y} = 26{,}7 + 0{,}00205x$$

em que

$\hat{y}$ = valor da compra
x = valor gasto nas lojas relacionadas

Usando essa equação, poderíamos prever que alguém, que gastou US$ 10 mil durante os últimos 39 meses nas lojas relacionadas, gastaria US$ 47,20 ao responder à promoção de mala direta. Neste capítulo, você aprenderá a desenvolver esse tipo de equação de regressão estimada.

O modelo final desenvolvido pelos analistas da ADS também incluiu diversas outras variáveis que aumentaram o poder de previsão da equação anterior. Algumas das variáveis incluídas foram a falta ou a posse de um cartão de crédito bancário, a renda estimada e a quantia média gasta em cada ida a uma loja selecionada. No próximo capítulo você aprenderá como se pode incorporar variáveis adicionais a um modelo de regressão múltipla.

As decisões administrativas freqüentemente se baseiam na relação entre duas ou mais variáveis. Por exemplo, depois de considerar a relação entre os gastos publicitários e as vendas, um gerente de marketing poderia tentar prever as vendas correspondentes a determinado nível de gastos publicitários. Em outro caso, uma empresa de serviços públicos poderia usar a relação entre a elevada temperatura diária e a demanda por eletricidade para prever o uso de energia elétrica com base na previsão de temperaturas elevadas para o próximo mês. Às vezes, o gerente pode recorrer à intuição para julgar a maneira pela qual duas variáveis estão relacionadas. Entretanto, se for possível obter os dados, um procedimento estatístico denominado *análise de regressão* pode ser usado para desenvolver uma equação que demonstra como as variáveis se relacionam.

Métodos estatísticos usados no estudo da relação entre duas variáveis foram empregados pela primeira vez por sir Francis Galton (1822-1911). Galton estava interessado em estudar a relação entre a altura de um pai e a altura de um filho. O discípulo de Galton, Karl Pearson (1857-1936), analisou a relação entre a altura de pais e filhos utilizando 1.078 pares de sujeitos.

Na terminologia da análise de regressão, a variável que é prevista é dita **variável dependente**. A variável ou variáveis usadas para prever o valor da variável dependente denominam-se **variáveis independentes**. Por exemplo, ao analisar o efeito dos gastos publicitários sobre as vendas, o desejo do gerente de marketing de prever as vendas sugeriria tornar as vendas a variável dependente. Os gastos publicitários seriam a variável independente usada para ajudar a prever as vendas. Na notação estatística, y designa a variável dependente e x, a variável independente.

Neste capítulo, consideraremos o tipo mais simples de análise de regressão envolvendo uma variável independente e uma variável dependente na qual a relação entre as variáveis se aproxima de uma linha reta. Ela é chamada **regressão linear simples**. A análise de regressão que envolve duas ou mais variáveis independentes denomina-se análise de regressão múltipla; a regressão múltipla será abordada no Capítulo 13.

12.1 MODELO DE REGRESSÃO LINEAR SIMPLES

A Armand's Pizza Parlors é uma rede de restaurantes de comida italiana localizada em cinco estados norte-americanos. As localizações mais bem-sucedidas dos restaurantes Armand's estão próximas a *campi* universitários. Os gerentes acreditam que as vendas trimestrais nesses restaurantes (designadas y) estão relacionadas positivamente com o tamanho da população estudantil (designado x); ou seja, os restaurantes próximos a *campi* universitários que contam com uma grande população estudantil tendem a gerar mais vendas que

aqueles que estão localizados próximos a *campi* que contam com uma pequena população estudantil. Usando análise de regressão, podemos determinar uma equação que mostra como a variável dependente y está relacionada com a variável independente x.

Modelo de Regressão e Equação de Regressão

No exemplo dos restaurantes Armand's Pizza Parlors, a população consiste em todos os restaurantes Armand's. Para cada restaurante da população, um valor x (população estudantil) corresponde a um valor y (vendas trimestrais). A equação que descreve como y está relacionado com x e com um termo de erro denomina-se **modelo de regressão.** O modelo de regressão usado na regressão linear simples é o seguinte:

MODELO DE REGRESSÃO LINEAR SIMPLES

$$y = \beta_0 + \beta_1 x + \epsilon \qquad (12.1)$$

$\beta_0 + \beta_1$ são chamados parâmetros do modelo, e ϵ (a letra grega epsílon) é uma variável aleatória que se denomina termo de erro. O termo de erro é responsável pela variabilidade em y que não pode ser explicada pela relação linear entre x e y.

A população de todos os restaurantes Armand's também pode ser vista como uma coleção de subpopulações, sendo uma para cada valor distinto de x. Por exemplo, uma subpopulação consiste em todos os restaurantes Armand's localizados próximo a *campi* universitários com 8 mil estudantes; outra subpopulação consiste em todos os restaurantes Armand's localizados próximo a *campi* universitários com 9 mil estudantes e assim por diante. Cada subpopulação tem uma distribuição correspondente de y valores. Desse modo, uma distribuição de y valores está associada a restaurantes localizados próximo a *campi* universitários com 8 mil estudantes; uma distribuição de y valores está associada a restaurantes localizados próximo a *campi* com 9 mil estudantes, e assim por diante. Cada distribuição de y valores tem sua própria média ou valor esperado. A equação que descreve como o valor esperado de y – designado por $E(y)$ – está relacionado com x denomina-se **equação de regressão.** A equação de regressão para a regressão linear simples é a seguinte:

EQUAÇÃO DE REGRESSÃO LINEAR SIMPLES

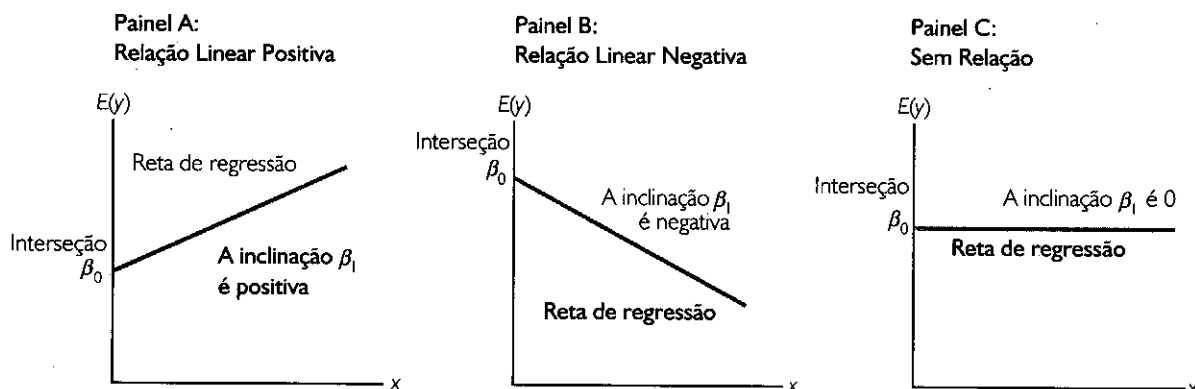
$$E(y) = \beta_0 + \beta_1 x \qquad (12.2)$$

O gráfico da equação de regressão linear simples é uma linha reta; β_0 é o ponto onde a linha (*ou* reta) de regressão intercepta o eixo y, β_1 é inclinação (declive) e $E(y)$ é a média ou valor esperado de y para determinado valor de x.

Exemplos de possíveis retas de regressão são mostrados na Figura 12.1. A reta de regressão do Painel A mostra que o valor médio de y está relacionado positivamente com x, e valores maiores de $E(y)$ estão associados a valores maiores de x. A reta de regressão do Painel B mostra que o valor médio de y está relacionado negativamente com x, e valores menores de $E(y)$ estão associados a valores maiores de x. A reta de regressão do Painel C apresenta o caso em que o valor médio de y não está relacionado com x; ou seja, o valor médio de y é o mesmo para todo valor de x.

Equação de Regressão Estimada

Se os valores dos parâmetros populacionais β_0 e β_1 fossem conhecidos, poderíamos usar a Equação 12.2 para calcular o valor médio de y para dado valor de x. Na prática, os valores paramétricos não são conhecidos, e precisam ser estimados usando-se os dados amostrais. A estatística da amostra (designada por b_0 e b_1) é calculada como estimativa dos parâmetros β_0 e β_1 da população. Substituindo os valores da estatística da amostra, b_0 e b_1 por β_0 e β_1 na equação de regressão, obtemos a **equação de regressão estimada.**

Figura 12.1 Retas de regressão possíveis na regressão linear simples

A equação de regressão estimada para a regressão linear simples é a seguinte:

EQUAÇÃO DE REGRESSÃO LINEAR SIMPLES ESTIMADA

$$\hat{y} = b_0 + b_1 x \qquad (12.3)$$

O gráfico da equação de regressão linear simples estimada denomina-se *reta de regressão ponderada*; b_0 é o ponto de interseção com o eixo y e b_1 é a inclinação. Na próxima seção, mostraremos como o método dos mínimos quadrados pode ser usado para calcular os valores de b_0 e b_1 na equação de regressão estimada.

Em geral, $\hat{y}$ é o estimador por ponto de $E(y)$, o valor médio de y para dado valor de x. Desse modo, para estimar a média ou o valor esperado das vendas trimestrais correspondentes a todos os restaurantes localizados nas proximidades de *campi* universitários com 10 mil estudantes, a Armand's substituiria o valor 10 mil por x na Equação 12.3. Em alguns casos, entretanto, a Armand's pode estar mais interessada em prever as vendas em um restaurante popular. Por exemplo, suponha que a Armand's queira prever as vendas trimestrais do restaurante localizado próximo ao Talbot College, uma escola com 10 mil estudantes. Ocorre que a melhor estimativa de y para dado valor de x também é fornecida por $\hat{y}$. Assim, para prever as vendas trimestrais no restaurante localizado próximo ao Talbot College, a Armand's substituiria o valor 10 mil por x na Equação 12.3.

Uma vez que o valor de $\hat{y}$ fornece uma estimação por ponto de $E(y)$ para determinado valor de x, tanto quanto uma estimação por ponto de um valor individual de y para dado valor de x, chamaremos $\hat{y}$ simplesmente de *valor estimado de y*. A Figura 12.2 apresenta um resumo do processo de estimação para regressão linear simples.

A estimação de β_0 e β_1 é um processo estatístico muito similar à estimação de μ discutida no Capítulo 7. β_0 e β_1 são os parâmetros de interesse desconhecidos, e b_0 e b_1 são as estatísticas amostrais usadas para estimar os parâmetros.

Figura 12.2 O processo de estimação em regressão linear simples

NOTAS E COMENTÁRIOS

1. A análise de regressão não pode ser interpretada como um procedimento para estabelecer uma relação de causa e efeito entre as variáveis. Ela somente é capaz de indicar como ou em que grau as variáveis estão associadas entre si. Quaisquer conclusões sobre causa e efeito devem basear-se no julgamento das pessoas que têm o melhor conhecimento da aplicação.
2. A equação de regressão da regressão linear simples é $E(y) = \beta_0 + \beta_1 x$. Livros mais avançados sobre análise de regressão freqüentemente grafam a equação de regressão como $E(y|x) = \beta_0 + \beta_1 x$ para enfatizar que a equação de regressão produz o valor médio de y para dado valor de x.

12.2 MÉTODO DOS MÍNIMOS QUADRADOS

Na regressão linear simples, cada observação consiste em dois valores: um para a variável independente e outro para a variável dependente.

O **método dos mínimos quadrados** é um procedimento que usa dados amostrais para encontrar a equação de regressão estimada. Para ilustrar o método dos mínimos quadrados, suponha que tenham sido coletados dados de uma amostra de dez restaurantes Armand's Pizza Parlors localizados nas proximidades de diversos *campi* universitários. Em relação à i-ésima observação ou restaurante da amostra, x_i é o tamanho da população estudantil (em milhares) e y_i são as vendas trimestrais (em milhares de dólares). Os valores de x_i e y_i correspondentes aos dez restaurantes da amostra estão resumidos na Tabela 12.1. Notamos que o restaurante 1, com $x_1 = 2$ e $y_1 = 58$, localiza-se próximo a um *campus* com 2 mil estudantes, e tem vendas trimestrais de US$ 58 mil. O restaurante 2, com $x_2 = 6$ e $y_2 = 105$, localiza-se próximo a um *campus* com 6 mil estudantes e tem vendas trimestrais de US$ 105 mil. O maior valor de vendas é do restaurante 10, que se localiza próximo a um *campus* com 26 mil estudantes e tem vendas trimestrais de US$ 202 mil.

A Figura 12.3 representa um diagrama de dispersão dos dados da Tabela 12.1. A população estudantil é mostrada no eixo horizontal e as vendas trimestrais, no eixo vertical. Os **diagramas de dispersão** para análise de regressão são construídos com a variável independente x no eixo horizontal e a variável dependente y no eixo vertical. O diagrama de dispersão nos possibilita observar os dados graficamente e tirar conclusões prévias sobre a possível relação entre as variáveis.

Quais conclusões prévias se pode tirar da Figura 12.3? As vendas trimestrais parecem ser mais elevadas nos *campi* que possuem maiores populações estudantis. Além disso, referente a esses dados, a relação entre o tamanho da população estudantil e as vendas trimestrais parece aproximar-se de uma linha reta; de fato, uma relação linear positiva é indicada entre x e y.

Tabela 12.1 Dados sobre a população de estudantes e as vendas trimestrais em dez restaurantes Armand's Pizza Parlors

Restaurante i	População de Estudantes (em milhares) x_i	Vendas Trimestrais (em milhares de dólares) y_i
1	2	58
2	6	105
3	8	88
4	8	118
5	12	117
6	16	137
7	20	157
8	20	169
9	22	149
10	26	202

ARQUIVO DA INTERNET
Armand's

Figura 12.3 Diagrama de dispersão da população de estudantes e das vendas trimestrais dos restaurantes Armand's Pizza Parlors

Por conseguinte, escolhemos o modelo de regressão linear simples para representar a relação entre as vendas trimestrais e a população de estudantes. Dada essa escolha, nossa próxima tarefa é usar os dados amostrais da Tabela 12.1 para determinar os valores de b_0 e b_1 na equação de regressão linear simples estimada. Para o i-ésimo restaurante, a equação de regressão estimada fornece:

$$\hat{y}_i = b_0 + b_1x_i \tag{12.4}$$

em que

$\hat{y}_i$ = valor estimado das vendas trimestrais (em milhares de dólares) para o i-ésimo restaurante
b_0 = o ponto em que a reta de regressão estimada intercepta y
b_1 = a inclinação da reta de regressão estimada
x_i = o tamanho da população estudantil (em milhares) para o i-ésimo restaurante

e y_i designa as vendas observadas (reais) do restaurante i e que $\hat{y}_i$ na Equação 12.4 representa o valor estimado das vendas do restaurante i, todo restaurante da amostra terá um valor observado de vendas y_i e um valor estimado de vendas $\hat{y}_i$. Para que a reta de regressão estimada produza um ajuste eficiente para os dados, queremos que as diferenças entre os valores de venda observados e os valores de venda estimados sejam pequenos.

O método dos mínimos quadrados utiliza dados amostrais para produzir os valores b_0 e b_1 que minimizam a *soma dos quadrados dos desvios* entre os valores observados da variável dependente y_i e os valores estimados da variável dependente. O critério utilizado no método dos mínimos quadrados é dado pela Equação 12.5.

Carl Friedrich Gauss (1777-1855) propôs o método dos mínimos quadrados.

CRITÉRIO DOS MÍNIMOS QUADRADOS

$$\min \Sigma(y_i - \hat{y}_i)^2 \tag{12.5}$$

em que

y_i = valor observado da variável dependente para a i-ésima observação
$\hat{y}_i$ = valor estimado da variável dependente para a i-ésima observação

Pode-se usar cálculo diferencial para mostrar que os valores de b_0 e de b_1 que minimizam a Equação 12.5 podem ser encontrados usando-se as Equações 12.6 e 12.7.

Para calcular b_1 com uma calculadora, carregue o maior número possível de dígitos significativos nos cálculos intermediários. Recomendamos carregar, no mínimo, quatro dígitos significativos.

INCLINAÇÃO E INTERSEÇÃO COM O EIXO y NA EQUAÇÃO DE REGRESSÃO ESTIMADA*

$$b_1 = \frac{\Sigma(x_i - \bar{x})(y_i - \bar{y})}{\Sigma(x_i - \bar{x})^2} \qquad \textbf{(12.6)}$$

$$b_0 = \bar{y} - b_1\bar{x} \qquad \textbf{(12.7)}$$

em que

x_i = valor da variável independente para a i-ésima observação
y_i = valor da variável dependente para a i-ésima observação
$\bar{x}$ = valor médio da variável independente
$\bar{y}$ = valor médio da variável dependente
n = número total de observações

Alguns dos cálculos necessários para se desenvolver a equação de regressão estimada pelo método dos mínimos quadrados para o caso dos restaurantes Armand's Pizza Parlors são mostrados na Tabela 12.2. Com a amostra de 10 restaurantes, temos $n = 10$ observações. Uma vez que as Equações 12.6 e 12.7 requerem $\bar{x}$ e $\bar{y}$, iniciamos os cálculos computando $\bar{x}$ e $\bar{y}$.

$$\bar{x} = \frac{\Sigma x_i}{n} = \frac{140}{10} = 14$$

$$\bar{y} = \frac{\Sigma y_i}{n} = \frac{1300}{10} = 130$$

Usando as Equações 12.6 e 12.7 e a informação contida na Tabela 12.2, podemos calcular a inclinação (declive) e a interseção da equação de regressão estimada referente ao exemplo dos restaurantes Armand's Pizza Parlors. O cálculo da inclinação (b_1) desenvolve-se da seguinte forma:

Tabela 12.2 Cálculos da equação de regressão estimada por mínimos quadrados para o caso dos restaurantes Armand's Pizza Parlors

Restaurante i	x_i	y_i	$x_i - \bar{x}$	$y_i - \bar{y}$	$(x_i - \bar{x})(y_i - \bar{y})$	$(x_i - \bar{x})^2$
1	2	58	212	272	864	144
2	6	105	28	225	200	64
3	8	88	26	242	252	36
4	8	118	26	212	72	36
5	12	117	22	213	26	4
6	16	137	2	7	14	4
7	20	157	6	27	162	36
8	20	169	6	39	234	36
9	22	149	8	19	152	64
10	26	202	12	72	864	144
Totais	140	1.300			2.840	568
	Σx_i	Σy_i			$\Sigma(x_i - \bar{x})(y_i - \bar{y})$	$\Sigma(x_i - \bar{x})^2$

* Uma fórmula alternativa para b_1 é

$$b_1 = \frac{\Sigma x_i y_i - (\Sigma x_i \Sigma y_i)/n}{\Sigma x_i^2 - (\Sigma x_i)^2/n}$$

Essa forma da Equação 12.6 freqüentemente é recomendada quando se usa uma calculadora para calcular b_1.

$$b_1 = \frac{\Sigma(x_i - \bar{x})(y_i - \bar{y})}{\Sigma(x_i - \bar{x})^2}$$

$$= \frac{2.840}{568}$$

$$= 5$$

O cálculo da interseção de y (b_0) é a seguinte:

Os Apêndices 12.1 e 12.2 mostram como o Minitab e o Excel podem ser usados para se obter a equação de regressão estimada.

$$b_0 = \bar{y} - b_1\bar{x}$$
$$= 130 - 5(14)$$
$$= 60$$

Desse modo, a equação de regressão estimada é

$$\hat{y} = 60 + 5x$$

A Figura 12.4 exibe o gráfico dessa equação no diagrama de dispersão.

A inclinação da equação de regressão estimada ($b_1 = 5$) é positiva, implicando que, à medida que a população estudantil aumenta, as vendas também sobem. Realmente, podemos concluir (com base nas vendas medidas em milhares de dólares e a população estudantil em milhares de alunos) que um aumento de mil alunos na população estudantil é associado a um acréscimo de US$ 5 mil nas vendas esperadas; ou seja, espera-se que as vendas trimestrais tenham um aumento de US$ 5 por estudante.

Se acreditarmos que a equação de regressão estimada pelo método dos mínimos quadrados descreve adequadamente a relação entre x e y, poderia parecer razoável usarmos a equação de regressão estimada para prever o valor de y para determinado valor de x. Por exemplo, se quiséssemos prever as vendas trimestrais de um restaurante a ser localizado próximo a um *campus* universitário com 16 mil estudantes, calcularíamos:

O uso da equação de regressão estimada para fazer previsões fora do intervalo de valores da variável independente deve ser feito com cautela porque, fora do intervalo, não podemos ter certeza de que a mesma relação é válida.

$$\hat{y} = 60 + 5(16) = 140$$

Portanto, preveríamos vendas trimestrais de US$ 140 mil para esse restaurante. Nas seções seguintes, discutiremos métodos para avaliar a conveniência de usar a equação de regressão estimada para fins de estimação e previsão.

Figura 12.4 Gráfico da equação de regressão estimada para os restaurantes Armand's Pizza Parlors: $\hat{y} = 60 + 5x$

NOTAS E COMENTÁRIOS

O método dos mínimos quadrados fornece uma equação de regressão estimada que minimiza a soma de desvios quadráticos entre os valores observados da variável dependente y_1 e os valores estimados da variável dependente $\hat{y}_i$. O critério dos mínimos quadrados é usado para escolher a equação que fornece o melhor ajuste. Se algum outro critério fosse usado, por exemplo, minimizar a soma dos desvios absolutos entre y_i e $\hat{y}_i$, uma equação diferente seria obtida. Na prática, o método dos mínimos quadrados é o mais amplamente usado.

Exercícios

Métodos

AUTOTESTE

1. São dadas cinco observações referentes a duas variáveis, x e y.

x_i	1	2	3	4	5
y_i	3	7	5	11	14

a. Desenvolva um diagrama de dispersão para esses dados.
b. O que o diagrama de dispersão desenvolvido no item (a) indica a respeito da relação entre as duas variáveis?
c. Tente aproximar a relação entre x e y traçando uma linha reta entre os dados.
d. Desenvolva a equação de regressão estimada calculando os valores de b_0 e b_1 usando as Equações 12.6 e 12.7.
e. Use a equação de regressão estimada para prever o valor de y quando $x = 4$.

2. São dadas cinco observações referentes a duas variáveis, x e y.

x_i	2	3	5	1	8
y_i	25	25	20	30	16

a. Desenvolva um diagrama de dispersão para esses dados.
b. O que o diagrama de dispersão desenvolvido no item (a) indica a respeito da relação entre as duas variáveis?
c. Tente aproximar a relação entre x e y traçando uma linha reta entre os dados.
d. Desenvolva a equação de regressão estimada calculando os valores de b_0 e b_1 usando as Equações 12.6 e 12.7.
e. Use a equação de regressão estimada para prever o valor de y quando $x = 6$.

3. São dadas cinco observações coletadas em um estudo de regressão sobre duas variáveis.

x_i	2	4	5	7	8
y_i	2	3	2	6	4

a. Desenvolva um diagrama de dispersão para esses dados.
b. Desenvolva a equação de regressão estimada para esses dados.
c. Use a equação de regressão estimada para prever o valor de y quando $x = 4$.

Aplicações

AUTOTESTE

4. Foram coletados os seguintes dados sobre altura (metros) e peso (quilogramas) de nadadoras:

Altura	1,72	1,63	1,57	1,65	1,68
Peso	59,87	48,98	46,26	52,16	58,05

a. Desenvolva um diagrama de dispersão desses dados, sendo a altura a variável independente.
b. O que o diagrama de dispersão desenvolvido no item (a) indica a respeito da relação entre as duas variáveis?
c. Tente aproximar a relação entre altura e peso traçando uma linha reta entre os dados.
d. Desenvolva a equação de regressão estimada calculando os valores de b_0 e b_1.
e. Se a altura de uma nadadora for 1,60 m, segundo sua estimativa, qual seria seu peso?

5. Os avanços tecnológicos ajudaram a tornar o *paddlecraft* inflável adequado para o uso em regiões distantes. Esses barcos infláveis de borracha, que podem ser enrolados em um feixe não muito maior que uma sacola de golfe, são grandes o bastante para acomodar um ou dois remadores e seus apetrechos

de *camping*. A revista *Canoe & Kayak* fez testes com barcos de nove fabricantes para determinar qual seria seu desempenho em uma viagem de três dias, a remo, por regiões ermas. Um dos critérios de avaliação foi a capacidade do barco para transportar bagagem, avaliada em uma escala de 4 pontos de 1 (a menor classificação) a 4 (a mais alta classificação). Os dados a seguir apresentam a avaliação da capacidade de transporte de bagagem e o preço do barco (*Canoe & Kayak*, março de 2003).

Barco	Capacidade de Transporte de Bagagem	Preço (US$)
S14	4	1.595
Orinoco	4	1.399
Outside Pro	4	1.890
Explorer 380X	3	795
River XK2	2,5	600
Sea Tiger	4	1.995
Maverik II	3	1.205
Starlite 100	2	583
Fat Pack Cat	3	1.048

a. Desenvolva um diagrama de dispersão desses dados, sendo a avaliação da capacidade de transporte de bagagem a variável independente.
b. O que o diagrama de dispersão desenvolvido no item (a) indica a respeito da relação entre capacidade e preço?
c. Trace uma linha reta entre os dados para fazer uma aproximação de uma relação linear entre a capacidade de transporte de bagagem e preço.
d. Use o método dos mínimos quadrados para desenvolver a equação de regressão estimada.
e. Apresente uma interpretação da inclinação da equação de regressão estimada.
f. Preveja o preço de um barco com capacidade de transporte de bagagem cuja avaliação é 3.

6. A Wageweb realiza pesquisas de dados salariais e apresenta os resumos em seu site. Com base nos dados salariais de 1º de outubro de 2002, a Wageweb divulgou que o salário médio anual dos vice-presidentes de vendas era US$ 142.111, com uma média de bonificação anual de US$ 15.432 (Wageweb.com, 13 de março de 2003). Suponha que os dados seguintes sejam uma amostra do salário anual e das bonificações de dez vice-presidentes de vendas. Os dados estão expressos em milhares de dólares.

Vice-Presidente	Salário	Bonificações
1	135	12
2	115	14
3	146	16
4	167	19
5	165	22
6	176	24
7	98	7
8	136	17
9	163	18
10	119	11

a. Desenvolva um diagrama de dispersão desses dados, sendo o salário a variável independente.
b. O que o diagrama de dispersão desenvolvido no item (a) indica a respeito da relação entre salários e bonificações?
c. Use o método dos mínimos quadrados para desenvolver a equação de regressão estimada.
d. Apresente uma interpretação da inclinação da equação de regressão estimada.
e. Preveja uma bonificação para um vice-presidente de vendas que recebe um salário anual de US$ 120 mil.

7. Você acha que os carros mais confiáveis custam mais caro? A *Consumer Reports* avaliou 15 sedãs de primeira-linha. A confiabilidade foi avaliada em uma escala de 5 pontos: fraca (1), razoável (2), boa (3), ótima (4) e excelente (5). As avaliações de preço e confiabilidade de cada um dos 15 carros são mostradas a seguir (*Consumer Reports*, fevereiro de 2004).

Marca e Modelo	Confiabilidade	Preço (US$)
Acura TL	4	33.150
BMW 330i	3	40.570
Lexus IS300	5	35.105
Lexus ES330	5	35.174
Mercedes-Benz C320	1	42.230
Lincoln LS Premium (V6)	3	38.225
Audi A4 3.0 Quattro	2	37.605
Cadillac CTS	1	37.695
Nissan Maxima 3.5 SE	4	34.390
Infiniti I35	5	33.845
Saab 9-3 Aero	3	36.910
Infiniti G35	4	34.695
Jaguar X-Type 3.0	1	37.995
Saab 9-5 Arc	3	36.955
Volvo S60 2.5T	3	33.890

ARQUIVO DA INTERNET

Cars

a. Desenvolva um diagrama de dispersão desses dados, sendo a avaliação da confiabilidade a variável independente.

b. Desenvolva a equação de regressão estimada por mínimos quadrados.

c. Com base em sua análise, você acha que os carros mais confiáveis custam mais caro? Explique.

d. Estime o preço de um sedã de primeira-linha que tenha uma avaliação de confiabilidade média.

8. *Mountain bikes* que custam menos de US$ 1.000 agora contêm muitos dos componentes de alta qualidade que até recentemente estavam disponíveis somente em modelos caros. Atualmente, até mesmo modelos que custam menos de US$ 1.000 muitas vezes oferecem suspensão elástica, *clipless pedals*[3] e estruturas altamente planejadas pela engenharia. Uma questão interessante é se o preço mais alto embute um nível mais elevado de manuseio, sendo este medido em termos da capacidade de *sidetrack*[4] da bicicleta. A *Outside Magazine* usou uma escala de classificação de 1 a 5, com 1 representando uma avaliação média e 5, uma avaliação excelente. A capacidade de *sidetrack* e o preço de dez bicicletas testadas pela *Outside Magazine* são apresentados a seguir (*Outside Magazine Buyer's Guide*, 2001).

Fábrica e Modelo	Capacidade de *Sidetrack*	Preço (US$)
Raleigh M80	1	600
Marin Bear Valley Feminina	1	649
GT Avalanche 2.0	2	799
Kona Jake the Snake	1	899
Schwinn Moab 2	3	950
Giant XTC NRS 3	4	1.100
Fisher Paragon Genesisters	4	1.149
Jamis Dakota XC	3	1.300
Trek Fuel 90	5	1.550
Specialized Stumpjumper M4	4	1.625

ARQUIVO DA INTERNET

MtnBikes

a. Desenvolva um diagrama de dispersão desses dados, sendo a capacidade de *sidetrack* a variável independente.

b. Parece que os modelos mais caros têm um nível de manuseio mais elevado? Explique.

c. Desenvolva a equação de regressão estimada por mínimos quadrados.

d. Qual é a estimativa de preço de uma *mountain bike* se ela tiver uma avaliação da capacidade de *sidetrack* igual a 4?

[3] NT: *Clipless pedal* – Tipo de pedal que contém um mecanismo que prende a sapatilha. Basta um movimento para liberar a sapatilha do pedal.

[4] NT: *Sidetrack* – Trilha ou caminho alternativo; terreno acidentado.

9. Um gerente de vendas coletou os seguintes dados sobre as vendas anuais e os anos de experiência profissional.

Vendedor	Anos de Experiência Profissional	Vendas Anuais (em milhares de dólares)
1	1	80
2	3	97
3	4	92
4	4	102
5	6	103
6	8	111
7	10	119
8	10	123
9	11	117
10	13	136

a. Desenvolva um diagrama de dispersão desses dados, sendo os anos de experiência profissional a variável independente.
b. Desenvolva uma equação de regressão estimada que poderia ser usada para prever as vendas anuais, dados os anos de experiência profissional.
c. Use a equação de regressão estimada para prever as vendas anuais efetuadas por um vendedor com nove anos de experiência profissional.

10. A *PC World* forneceu avaliações dos 15 melhores computadores *notebook* (*PC World*, fevereiro de 2000). A avaliação do desempenho é uma medida de quão rapidamente um PC é capaz de rodar uma combinação de aplicativos comerciais comuns em comparação com o desempenho de velocidade de uma máquina básica para realizar a mesma tarefa. Por exemplo, um PC com uma avaliação de desempenho igual a 200 é duas vezes mais rápido que a máquina básica. Foi usada uma escala de 100 pontos para representar a classificação geral de cada *notebook* testado no estudo. Uma avaliação na faixa dos 90 pontos é excepcional, ao passo que uma avaliação na faixa dos 70 pontos está acima da média. As avaliações de desempenho e as classificações gerais dos 15 *notebooks* são as seguintes:

ARQUIVO DA INTERNET PCs

Marca e Modelo	Avaliação do Desempenho	Classificação Geral
AMS Tech Roadster 15CTA380	115	67
Compaq Armada M700	191	78
Compaq Prosignia Notebook 150	153	79
Dell Inspiron 3700 C466GT	194	80
Dell Inspiron 7500 R500VT	236	84
Dell Latitude Cpi A366XT	184	76
Enpower ENP-313 Pro	184	77
Gateway Solo 9300LS	216	92
HP Pavilion Notebook PC	185	83
IBM ThinkPad I Series 1480	183	78
Micro Express NP7400	189	77
Micron TransPort NX PII-400	202	78
NEC Versa SX	192	78
Sceptre Soundx 5200	141	73
Sony VAIO PCG-F340	187	77

a. Desenvolva um diagrama de dispersão desses dados, sendo a avaliação de desempenho a variável independente.
b. Desenvolva a equação de regressão estimada por mínimos quadrados.
c. Estime a classificação global do novo PC que tem uma pontuação de desempenho igual a 225.

11. Não obstante os atrasos nos grandes aeroportos agora serem menos freqüentes, é útil saber quais aeroportos têm probabilidade de fazê-lo perder o horário de seus compromissos. Além disso, se o seu avião chegar atrasado em um aeroporto em particular onde você deve fazer uma conexão, qual é a probabilidade de a partida se atrasar e, dessa forma, aumentar suas chances de fazer a conexão? Os dados a seguir mostram a porcentagem de chegadas e partidas atrasadas durante o mês de agosto em 13 aeroportos (*Business 2.0*, fevereiro de 2002).

Aeroporto	Chegadas Atrasadas (%)	Partidas Atrasadas (%)
Atlanta	24	22
Charlotte	20	20
Chicago	30	29
Cincinnati	20	19
Dallas	20	22
Denver	23	23
Detroit	18	19
Houston	20	16
Minneapolis	18	18
Phoenix	21	22
Pittsburgh	25	22
Salt Lake City	18	17
St. Louis	16	16

ARQUIVO DA INTERNET
Airport

a. Desenvolva um diagrama de dispersão desses dados, sendo a porcentagem de chegadas atrasadas a variável independente.
b. O que o diagrama de dispersão desenvolvido no item (a) indica a respeito da relação entre chegadas atrasadas e partidas atrasadas?
c. Use o método dos mínimos quadrados para desenvolver a equação de regressão estimada.
d. Apresente uma interpretação da inclinação da equação de regressão estimada.
e. Suponha que a porcentagem de chegadas atrasadas no aeroporto de Filadélfia durante o mês de agosto tenha sido 22%. Qual é a estimativa da porcentagem de partidas atrasadas?

12. A tabela seguinte apresenta o número de empregados e a receita (em milhões de dólares) de 20 empresas (*Fortune*, 17 de abril de 2000).

Empresa	Empregados	Receita (milhões de dólares)
Sprint	77.600	19.930
Chase Manhattan	74.801	33.710
Computer Sciences	50.000	7.660
Wells Fargo	89.355	21.795
Sunbeam	12.200	2.398
CBS	29.000	7.510
Time Warner	69.722	27.333
Steelcase	16.200	2.743
Georgia-Pacific	57.000	17.796
Toro	1.275	4.673
American Financial	9.400	3.334
Fluor	53.561	12.417
Phillips Petroleum	15.900	13.852
Cardinal Health	36.000	25.034
Borders Group	23.500	2.999
MCI Worldcom	77.000	37.120
Consolidated Edison	14.269	7.491
IBP	45.000	14.075
Super Value	50.000	17.421
H&R Block	4.200	1.669

ARQUIVO DA INTERNET
EmpRev

a. Desenvolva um diagrama de dispersão desses dados, sendo o número de empregados a variável independente.
b. O que o diagrama de dispersão desenvolvido no item (a) indica a respeito da relação entre o número de empregados e a receita?
c. Desenvolva a equação de regressão estimada desses dados.
d. Use a equação de regressão estimada para prever a receita de uma empresa com 75 mil empregados.

13. Para o Internal Revenue Service (Departamento da Receita Federal), a aceitabilidade das deduções totais detalhadas depende da renda bruta ajustada do contribuinte. Deduções vultosas, que incluem doações a obras assistenciais e deduções de despesas médicas, são mais aceitáveis para contribuintes que têm grandes rendas brutas ajustadas. Se um contribuinte reivindicar um valor maior que a média das deduções detalhadas para determinado nível de renda, elevam-se as probabilidades de uma auditoria do IRS. Dados (em milhares de dólares) sobre a renda bruta ajustada e a média, ou valor aceitável, das deduções detalhadas são apresentados a seguir:

Renda Bruta Ajustada (milhares de dólares)	Valor Aceitável de Deduções Detalhadas (milhares de dólares)
22	9,6
27	9,6
32	10,1
48	11,1
65	13,5
85	17,7
120	25,5

a. Desenvolva um diagrama de dispersão desses dados, sendo a renda bruta ajustada a variável independente.
b. Use o método dos mínimos quadrados para desenvolver a equação de regressão estimada.
c. Faça uma estimativa do nível aceitável de deduções detalhadas para um contribuinte que tem uma renda bruta ajustada de US$ 52.500. Se esse contribuinte reivindicasse deduções detalhadas de US$ 20.400, um pedido de auditoria por um fiscal do IRS pareceria justificável? Explique.

14. Os salários iniciais dos contadores e auditores de Rochester, NY, acompanham os de muitas cidades dos Estados Unidos. Os dados a seguir apresentam o salário inicial (em milhares de dólares) e o índice do custo de vida de Rochester e de nove outras regiões metropolitanas (*Democrat and Chronicle*, 1º de setembro de 2002). O índice do custo de vida, baseado no preço dos alimentos, moradia, impostos e outros custos, varia de 0 (o mais caro) a 100 (o mais barato).

ARQUIVO DA INTERNET
Salaries

Região Metropolitana	Índice	Salário (US$1.000)
Oklahoma City	82,44	23,9
Tampa/St, Petersburg/Clearwater	79,89	24,5
Indianapolis	55,53	27,4
Buffalo/Niagara Falls	41,36	27,7
Atlanta	39,38	27,1
Rochester	28,05	25,6
Sacramento	25,50	28,7
Raleigh/Durham/Chapel Hill	13,32	26,7
San Diego	3,12	27,8
Honolulu	0,57	28,3

a. Desenvolva um diagrama de dispersão desses dados, sendo o índice do custo de vida a variável independente.
b. Desenvolva a equação de regressão estimada relacionando o índice do custo de vida com o salário inicial.
c. Estime o salário inicial de uma região metropolitana que tem um índice do custo de vida igual a 50.

12.3 COEFICIENTE DE DETERMINAÇÃO

No exemplo dos restaurantes Armand's Pizza Parlors, desenvolvemos a equação de regressão estimada $\hat{y} = 60 + 5x$ para aproximar a relação linear entre o tamanho da população estudantil x e as vendas trimestrais y. A questão agora é: quão satisfatoriamente a equação de regressão estimada ajusta os dados? Nesta seção, mostraremos que o **coeficiente de determinação** nos dá uma medida da eficiência de ajuste da equação de regressão estimada.

Em relação à i-ésima observação, a diferença entre o valor observado da variável dependente, y_i, e o valor estimado da variável dependente, $\hat{y}_i$, denomina-se ***i*-ésimo resíduo**. O i-ésimo resíduo representa o erro de usarmos $\hat{y}_i$ para estimar y_i. Dessa forma, para a i-ésima observação, o resíduo é $y_i - \hat{y}_i$. A soma dos quadrados desses resíduos ou erros é a quantidade que é minimizada pelo método dos mínimos quadrados. Essa quantidade, também conhecida como a soma dos quadrados dos erros (*sum of squares due to error*), é designada por SSE.

SOMA DOS QUADRADOS DOS ERROS

$$SSE = \Sigma(y_i - \hat{y}_i)^2 \qquad \textbf{(12.8)}$$

O valor da SSE é uma medida do erro de se usar a equação de regressão estimada para estimar os valores da variável dependente na amostra.

Na Tabela 12.3, apresentamos os cálculos necessários para se obter a soma dos quadrados dos erros para o exemplo dos restaurantes Armand's Pizza Parlors. Por exemplo, para o restaurante 1, os valores das variáveis independente e dependente são $x_1 = 2$ e $y_1 = 58$. Usando a equação de regressão estimada, descobrimos que o valor estimado das vendas trimestrais para o restaurante 1 é $\hat{y}_1 = 60 + 5(2) = 70$. Desse modo, o erro de se usar $\hat{y}_1$ para estimar y_1 para o restaurante 1 é $y_1 - \hat{y}_1 = 58 - 70 = -12$. O erro elevado ao quadrado, $(-12)^2 = 144$, é mostrado na última coluna da Tabela 12.3. Depois de calcular e elevar ao quadrado os resíduos correspondentes a cada restaurante da amostra, fazemos seu somatório e obtemos SSE = 1.530. Assim, SSE = 1.530 mede o erro de se usar a equação de regressão estimada $\hat{y}_1 = 60 + 5x$ para prever as vendas.

Suponha agora que nos peçam para desenvolver uma estimativa das vendas trimestrais sem sabermos qual é o tamanho da população estudantil. Sem ter o conhecimento de nenhuma das variáveis relacionadas, usaríamos a média amostral como uma estimativa das vendas trimestrais em qualquer restaurante dado.

Tabela 12.3 Cálculo da SSE para os restaurantes Armand's Pizza Parlors

Restaurante i	x_i = População Estudantil (em milhares)	y_i = Vendas Trimestrais (em milhares de dólares)	Vendas Previstas $\hat{y}_i = 60 + 5x_i$	Erro $y_i - \hat{y}_i$	Erro Elevado ao Quadrado $(y_i - \hat{y}_i)^2$
1	2	58	70	−12	144
2	6	105	90	15	225
3	8	88	100	−12	144
4	8	118	100	18	324
5	12	117	120	−3	9
6	16	137	140	−3	9
7	20	157	160	−3	9
8	20	169	160	9	81
9	22	149	170	−21	441
10	26	202	190	12	144
					SSE = 1.530

Tabela 12.4 Cálculo da soma total dos quadrados para os restaurantes Armand's Pizza Parlors

Restaurante i	x_i = População Estudantil (em milhares)	y_i = Vendas Trimestrais (em milhares de dólares)	Desvio $y_i - \bar{y}$	Desvio Elevado ao Quadrado $(y_i - \bar{y})^2$
1	2	58	−72	5.184
2	6	105	−25	625
3	8	88	−42	1.764
4	8	118	−12	144
5	12	117	−13	169
6	16	137	7	49
7	20	157	27	729
8	20	169	39	1.521
9	22	149	19	361
10	26	202	72	5.184
				SST = 15.730

A Tabela 12.2 indica que, para os dados de vendas, $\Sigma y_i = 1.300$. Portanto, o valor médio das vendas trimestrais para a amostra de dez restaurantes Armand's é $\bar{y} = \Sigma y_i/n = 1.300/10 = 130$. Na Tabela 12.4, mostramos a soma dos desvios elevados ao quadrado, obtida usando-se a média amostral $\bar{y} = 130$ para estimar o valor das vendas trimestrais correspondentes a cada restaurante da amostra. Em relação ao i-ésimo restaurante da amostra, a diferença $y_i - \bar{y}$ fornece a medida do erro envolvido no uso de $\bar{y}$ para estimar as vendas. A soma dos quadrados correspondente, denominada *soma total dos quadrados* (*total sum of squares*), é designada por SST.

SOMA TOTAL DOS QUADRADOS

$$SST = \Sigma(y_i - \bar{y})^2 \qquad (12.9)$$

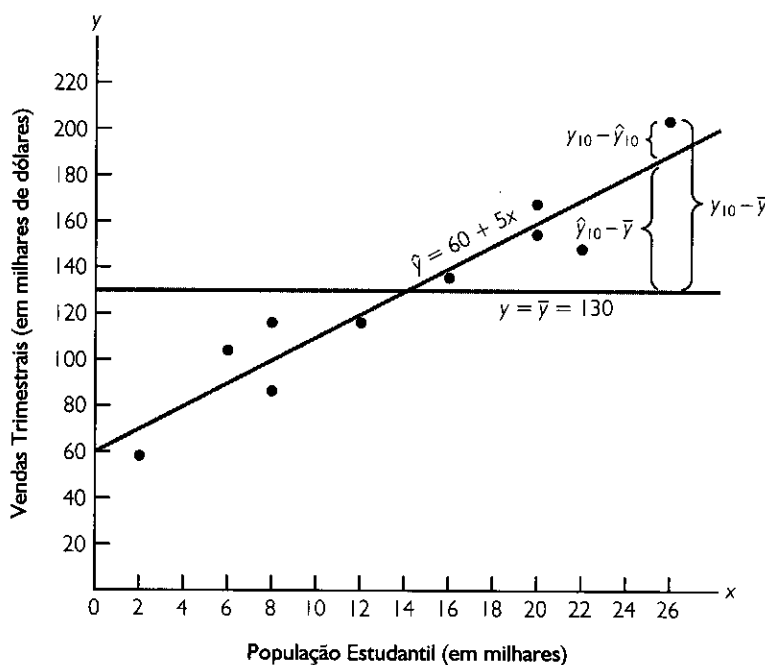
A soma na parte inferior da última coluna da Tabela 12.4 é a soma total dos quadrados do exemplo dos restaurantes Armand's Pizza Parlors; ou seja, SST = 15.730.

Na Figura 12.5, mostramos a reta de regressão estimada $\hat{y} = 60 + 5x$ e a reta correspondente a $\bar{y} = 130$. Observe que os pontos se agrupam mais estreitamente em torno da reta de regressão estimada do que nas proximidades da reta $\bar{y} = 130$. Por exemplo, em relação ao décimo restaurante da amostra, notamos que o erro é muito maior quando $\bar{y} = 130$ é usado como uma estimativa de y_{10} do que quando $\hat{y}_{10} = 60 + 5(26) = 190$ é usado. Podemos imaginar a SST como uma medida de quão satisfatoriamente as observações se agrupam nas proximidades na reta $\hat{y}$.

Para medir quanto os valores de $\hat{y}$ na reta de regressão estimada se afastam de $\bar{y}$, outra soma de quadrados é calculada. Essa soma de quadrados, denominada *soma dos quadrados da regressão* (*sum of squares due to regression*), é designada SSR.

SOMA DOS QUADRADOS DA REGRESSÃO

$$SSR = \Sigma(\hat{y}_i - \bar{y})^2 \qquad (12.10)$$

Figura 12.5 Desvios nas proximidades reta de regressão estimada e da reta $y = \bar{y}$ para os restaurantes Armand's Pizza Parlors

Da discussão anterior, devemos esperar que SST, SSR e SSE estejam relacionadas. De fato, a relação entre essas três somas de quadrados produz um dos resultados mais importantes em estatística.

A SSR pode ser imaginada como a parte explicada da SST, e a SSE pode ser imaginada como a parte não explicada da SST.

RELAÇÃO ENTRE SST, SSR E SSE

$$SST = SSR + SSE \qquad (12.11)$$

em que

SST = soma total dos quadrados
SSR = soma dos quadrados da regressão
SSE = soma dos quadrados dos erros

A Equação 12.11 mostra que a soma total dos quadrados pode ser dividida em dois componentes: a soma dos quadrados da regressão e a soma dos quadrados dos erros. Portanto, se os valores de duas quaisquer dessas somas de quadrados forem conhecidos, a terceira soma de quadrados poderá ser calculada facilmente. Por exemplo, no caso dos restaurantes Armand's Pizza Parlors, já sabemos que SSE = 1.530 e que SST = 15.730; portanto, resolvendo SSR na Equação 12.11, descobrimos que a soma dos quadrados da regressão é:

$$SSR = SST - SSE = 15.730 - 1.530 = 14.200$$

Vejamos agora como as três somas de quadrados, SST, SSR e SSE, podem ser usadas para fornecer uma medida da eficiência de ajuste da equação de regressão estimada. A equação de regressão estimada forneceria um ajuste perfeito se todo valor da variável dependente y_i se situasse na reta de regressão estimada. Nesse caso, $y_i - \hat{y}_i$ seria igual a zero para cada observação, resultando em SSE = 0. Uma vez que SST = SSR + SSE, notamos que para haver um ajuste perfeito SSR deve igualar-se a SST, e a razão (SSR/SST) deve ser igual a 1. Ajustes mais imperfeitos resultarão em valores maiores para SSE. Resolvendo SSE na Equação 12.11, notamos que SSE = SST – SSR. Portanto, o maior valor para SSE (e, daí, o pior ajuste) ocorre quando SSR = 0 e SSE = SST.

A razão SSR/SST, a qual assumirá valores entre zero e 1, é usada para avaliar a eficiência de ajuste da equação de regressão estimada. Essa razão é chamada *coeficiente de determinação* e é designada por r^2.

COEFICIENTE DE DETERMINAÇÃO

$$r^2 = \frac{SSR}{SST} \qquad (12.12)$$

Para o exemplo dos restaurantes Armand's Pizza Parlors, o valor do coeficiente de determinação é:

$$r^2 = \frac{SSR}{SST} = \frac{14.200}{15.730} = 0{,}9027$$

Quando expressamos o coeficiente de determinação como uma porcentagem, r^2 pode ser interpretado como a porcentagem da soma total dos quadrados que pode ser explicada usando-se a equação de regressão estimada. Em relação aos restaurantes Armand's Pizza Parlors, podemos concluir que 90,27% da soma total dos quadrados pode ser explicada usando-se a equação de regressão estimada $\hat{y} = 60 + 5x$ para prever as vendas trimestrais. Em outras palavras, 90,27% da variabilidade das vendas podem ser explicados por meio da relação linear existente entre o tamanho da população estudantil e as vendas. Ficaríamos satisfeitos em encontrar um ajuste tão bom para a equação de regressão estimada.

Coeficiente de Correlação

No Capítulo 3, apresentamos o **coeficiente de correlação** como uma medida da intensidade da associação linear entre duas variáveis, x e y. Os valores do coeficiente de correlação estão sempre entre –1 e +1. Um valor +1 indica que as duas variáveis x e y estão perfeitamente relacionadas em um sentido linear positivo. Ou seja, todos os pontos de dados estão em uma linha reta que tem uma inclinação positiva. Um valor –1 indica que x e y estão perfeitamente relacionadas em um sentido linear negativo, com todos os pontos de dados em uma linha reta que tem uma inclinação negativa. Valores do coeficiente de correlação próximos de zero indicam que x e y não estão linearmente relacionadas.

Na Seção 3.5, apresentamos a equação para calcular o coeficiente de correlação da amostra. Se uma análise de regressão já tiver sido realizada e o coeficiente de determinação r^2, calculado, o coeficiente de correlação da amostra pode ser calculado da seguinte maneira:

COEFICIENTE DE CORRELAÇÃO DA AMOSTRA

$$\begin{aligned} r_{xy} &= (\text{sinal de } b_1)\sqrt{\text{Coeficiente de determinação}} \\ &= (\text{sinal de } b_1)\sqrt{r^2} \end{aligned} \qquad (12.13)$$

em que

$$b_1 = \text{a inclinação da equação de regressão estimada } \hat{y} = b_0 + b_1 x$$

O sinal do coeficiente de correlação da amostra será positivo se a equação de regressão estimada tiver uma inclinação positiva ($b_1 > 0$) e será negativo se a equação de regressão estimada tiver uma inclinação negativa ($b_1 < 0$).

Em relação ao exemplo dos restaurantes Armand's Pizza Parlors, o valor do coeficiente de determinação correspondente à equação de regressão estimada $\hat{y} = 60 + 5x$ é 0,9027. Uma vez que a inclinação da equação de regressão estimada é positiva, a Equação (12.13) mostra que o coeficiente de correlação da amostra é $+\sqrt{0,9027} = +0,9501$. Com um coeficiente de correlação da amostra igual a $r_{xy} = +0,9501$, concluiríamos que existe uma forte associação linear positiva entre x e y.

No caso de uma relação linear entre duas variáveis, tanto o coeficiente de determinação como o coeficiente de correlação da amostra fornecem medidas da intensidade da relação. O coeficiente de determinação fornece uma medida entre zero e 1, ao passo que o coeficiente de correlação da amostra fornece uma medida entre –1 e +1. Embora o coeficiente de correlação da amostra se restrinja a uma relação linear entre duas variáveis, o coeficiente de determinação pode ser usado para relações não lineares e para relações que têm duas ou mais variáveis independentes. Desse modo, o coeficiente de determinação fornece uma faixa mais ampla de aplicabilidade.

NOTAS E COMENTÁRIOS

1. Para desenvolver a equação de regressão estimada pelo método dos mínimos quadrados e calcular o coeficiente de determinação, não fizemos suposições probabilísticas à respeito do termo de erro **e** e não foi realizado nenhum teste estatístico quanto à significância da relação entre x e y. Valores maiores de r^2 implicam que a reta dos mínimos quadrados provê melhor ajuste para os dados; ou seja, as observações se agrupam mais estreitamente nas proximidades da reta dos mínimos quadrados. Mas, usando somente r^2, não podemos tirar nenhuma conclusão a respeito de a relação entre x e y ser ou não ser estatisticamente significativa. Essa conclusão deve basear-se em considerações que envolvem o tamanho da amostra e as propriedades da distribuição amostral apropriada dos estimadores mínimos quadrados.
2. Na prática, para dados típicos encontrados nas ciências sociais, valores de r^2 pequenos, de até 0,25, muitas vezes são considerados úteis. Para dados das ciências físicas e das ciências biológicas, valores de r^2 iguais a 0,60 ou maiores freqüentemente são considerados úteis. Realmente, em alguns casos, valores de r^2 superiores a 0,90 podem ser encontrados. Em aplicações de negócios, os valores de r^2 variam muito, dependendo das características particulares a cada aplicação.

Exercícios

Métodos

AUTOTESTE

15. Os dados do exercício 1 são os seguintes:

x_i	1	2	3	4	5
y_i	3	7	5	11	14

A equação de regressão estimada para esses dados é $\hat{y} = 0,20 + 2,60x$.

a. Calcule SSE, SST e SSR usando as Equações 12.8, 12.9 e 12.10.
b. Calcule o coeficiente de determinação r^2. Comente a eficiência de ajuste.
c. Calcule o coeficiente de correlação da amostra.

16. Os dados do exercício 2 são os seguintes:

x_i	2	3	5	1	8
y_i	25	25	20	30	16

A equação de regressão estimada para esses dados é $\hat{y} = 30,33 - 1,88x$.

a. Calcule SSE, SST e SSR.
b. Calcule o coeficiente de determinação r^2. Comente a eficiência de ajuste.
c. Calcule o coeficiente de correlação da amostra.

17. Os dados do exercício 3 são os seguintes:

x_i	2	4	5	7	8
y_i	2	3	2	6	4

A equação de regressão estimada para esses dados é $\hat{y} = 0{,}75 + 0{,}51x$. Qual porcentagem da soma total dos quadrados pode ser levada em conta pela equação de regressão estimada? Qual é o valor do coeficiente de correlação da amostra?

Aplicações

AUTOTESTE

18. Os dados a seguir são os salários mensais y e o – *grade point average*[5] (GPA) – y de estudantes que obtiveram o diploma de bacharel em administração comercial com habilitação em sistemas de informação. A equação de regressão estimada para esses dados é $\hat{y} = 1.790{,}5 + 581{,}1x$.

GPA	Salário Mensal (US$)	GPA	Salário Mensal (US$)
2,6	3.300	3,2	3.500
3,4	3.600	3,5	3.900
3,6	4.000	2,9	3.600

a. Calcule SST, SSR e SSE.
b. Calcule o coeficiente de determinação r^2. Comente a eficiência de ajuste.
c. Qual é o valor do coeficiente de correlação da amostra?

19. Os dados do exercício 7 são os seguintes:

ARQUIVO DA INTERNET
Cars

Marca e Modelo	x = Confiabilidade	y = Preço (US$)
Acura TL	4	33.150
BMW 330i	3	40.570
Lexus IS300	5	35.105
Lexus ES330	5	35.174
Mercedes-Benz C320	1	42.230
Lincoln LS Premium (V6)	3	38.225
Audi A4 3.0 Quattro	2	37.605
Cadillac CTS	1	37.695
Nissan Maxima 3.5 SE	4	34.390
Infiniti I35	5	33.845
Saab 9-3 Aero	3	36.910
Infiniti G35	4	34.695
Jaguar X-Type 3.0	1	37.995
Saab 9-5 Arc	3	36.955
Volvo S60 2.5T	3	33.890

A equação de regressão estimada desses dados é $\hat{y} = 40.639 - 1.301x$. Qual porcentagem da soma total de quadrados pode ser levada em conta pela equação de regressão estimada? Comente a eficiência de ajuste. Qual é o coeficiente de correlação da amostra?

20. A renda familiar típica e o preço típico das moradias referentes a uma amostra de 18 cidades são os seguintes (*Places Rated Almanac*, 2000). Os dados estão expressos em milhares de dólares.

ARQUIVO DA INTERNET
Cities

Cidade	Renda	Preço das casas
Akron, OH	74,1	114,9
Atlanta, GA	82,4	126,9
Birmingham, AL	71,2	130,9
Bismarck, ND	62,8	92,8
Cleveland, OH	79,2	135,8
Columbia, SC	66,8	116,7
Denver, CO	82,6	161,9
Detroit, MI	85,3	145,0
Fort Lauderdale, FL	75,8	145,3
Hartford, CT	89,1	162,1
Lancaster, PA	75,2	125,9
Madison, WI	78,8	145,2
Naples, FL	100,0	173,6

[5] NT: *Grade Point Average* (GPA) – Média de notas, média escolar. Uma medida numérica do rendimento acadêmico baseada no cálculo do número de créditos e notas obtidas em todas as matérias. Baseia-se em uma escala de 0 a 4.

Cidade	Renda	Preço das Casas
Nashville, TN	77,3	125,9
Philadelphia, PA	87,0	151,5
Savannah, GA	67,8	108,1
Toledo, OH	71,2	101,1
Washington, DC	97,4	191,9

a. Com esses dados, desenvolva uma equação de regressão estimada que possa ser usada para estimar o preço típico das moradias de uma cidade, dada a renda familiar típica.
b. Calcule r^2. Você se sentiria à vontade em usar essa equação de regressão estimada para estimar o preço típico de uma moradia em uma cidade?
c. Estime a preço típico de uma moradia em uma cidade que tem a renda familiar típica de US$ 95 mil.

21. Uma importante aplicação da análise de regressão na contabilidade é a estimação do custo. Ao coletar dados sobre volume e custo e usar o método dos mínimos quadrados para desenvolver uma equação de regressão estimada relacionando volume e custo, um contador pode estimar o custo associado a um volume de manufatura em particular. Considere a seguinte amostra de volumes de produção e os dados de custos totais referentes a uma operação de manufatura.

Volume de Produção (unidades)	Custos Totais (US$)
400	4.000
450	5.000
550	5.400
600	5.900
700	6.400
750	7.000

a. Com esses dados, desenvolva uma equação de regressão estimada que possa ser usada para prever o custo total de determinado volume de produção.
b. Qual é o custo variável por unidade produzida?
c. Calcule o coeficiente de determinação. Qual porcentagem da variação no custo total pode ser explicada pelo volume de produção?
d. O programa de produção da empresa mostra que 500 unidades devem ser produzidas no próximo mês. Qual é o custo total estimado para essa operação?

22. A revista *PC World* divulgou avaliações das cinco melhores impressoras a laser para pequenos escritórios e cinco impressoras a laser para corporações (*PC World*, fevereiro de 2003). A impressora a laser para pequenos escritórios mais bem classificada foi a Minolta-QMS PagePro 1250W, com uma avaliação global igual a 91. A impressora a laser para corporações mais bem classificada, a Xerox Phaser 4400/N, obteve uma avaliação global igual a 83. Os dados a seguir revelam a velocidade de impressão de texto simples em termos de páginas por minuto (ppm) e o preço de cada impressora.

ARQUIVO DA INTERNET
Printers

Marca	Tipo	Velocidade(ppm)	Preço ($)
Minolta-QMS PagePro 1250W	Pequeno Escritório	12	199
Brother HL-1850	Pequeno Escritório	10	499
Lexmark E320	Pequeno Escritório	12,2	299
Minolta-QMS PagePro 1250E	Pequeno Escritório	10,3	299
HP Laserjet 1200	Pequeno Escritório	11,7	399
Xerox Phaser 4400/N	Corporativa	17,8	1.850
Brother HL-2460N	Corporativa	16,1	1.000
IBM Infoprint 1120n	Corporativa	11,8	1.387
Lexmark W812	Corporativa	19,8	2.089
Oki Data B8300n	Corporativa	28,2	2.200

a. Desenvolva a equação de regressão estimada, sendo a velocidade a variável independente.
b. Calcule r^2. Qual porcentagem da variação de custo pode ser explicada pela velocidade da impressora?
c. Qual é o coeficiente de correlação amostral entre velocidade e preço? Ele reflete uma relação forte ou fraca entre a velocidade de impressão e o custo?

12.4 SUPOSIÇÕES DO MODELO

Ao realizar uma análise de regressão, inicie fazendo uma suposição sobre o modelo apropriado para a relação entre a(s) variável(is) dependente(s) e independente(s). Para o caso de uma regressão linear simples, o modelo de regressão suposto é:

$$y = \beta_0 + \beta_1 x + \epsilon$$

Então, o método dos mínimos quadrados é usado para desenvolver valores para b_0 e b_1, os quais são as estimativas dos parâmetros modelo β_0 e β_1, respectivamente. A equação de regressão estimada resultante é:

$$\hat{y} = b_0 + b_1 x$$

Vimos que o valor do coeficiente de determinação (r^2) é uma medida da eficiência de ajuste da equação de regressão estimada. Entretanto, mesmo com um valor grande de r^2, a equação de regressão estimada não deve ser usada enquanto não se fizer uma análise adicional da adequabilidade do modelo suposto. Uma etapa importante para determinar se o modelo suposto é apropriado envolve testar a significância da relação. Os testes de significância na análise de regressão baseiam-se nas seguintes suposições sobre o termo de erro ϵ.

SUPOSIÇÕES SOBRE O TERMO DE ERRO ϵ NO MODELO DE REGRESSÃO

$$y = \beta_0 + \beta_1 x + \epsilon$$

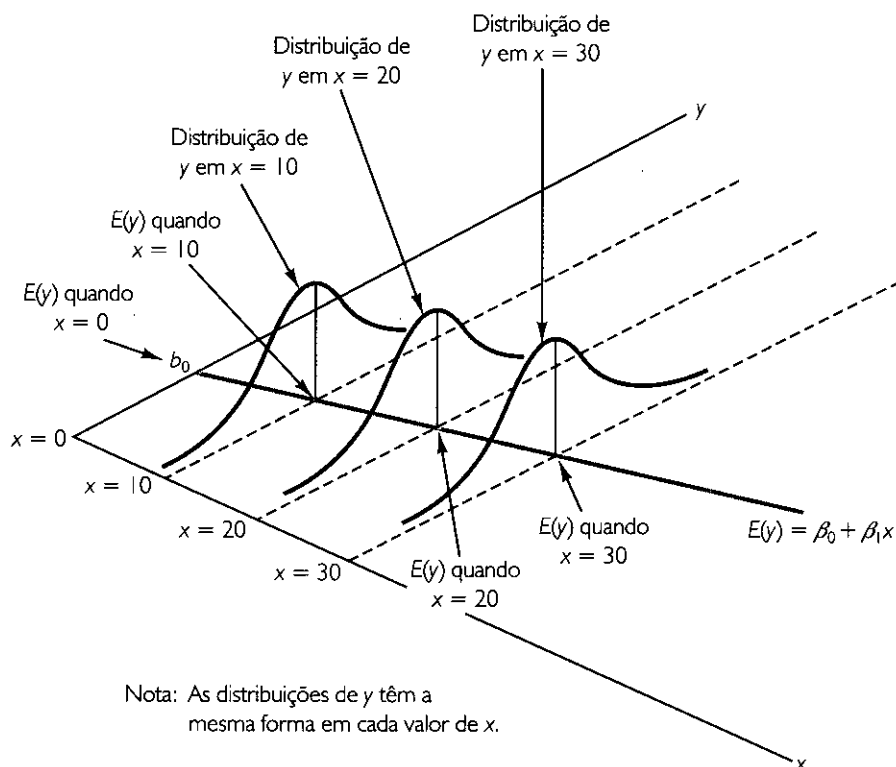
1. O termo de erro e é uma variável aleatória com uma média, ou valor esperado, igual a zero; ou seja, $E(\epsilon) = 0$.
   Implicação: β_0 e β_1 são constantes; por conseguinte, $E(\beta_0) = \beta_0$ e $E(\beta_1) = \beta_1$; desse modo, para dado valor de x, o valor esperado de y é

   $$E(y) = \beta_0 + \beta_1 x \quad \textbf{(12.14)}$$

   Conforme indicamos anteriormente, a Equação 12.14 é chamada equação de regressão.
2. A variância de ϵ, designada por σ^2, é a mesma para todos os valores de x.
   Implicação: A variância de y nas proximidades da reta de regressão é igual a σ^2 e é a mesma para todos os valores de x.
3. Os valores de ϵ são independentes.
   Implicação: O valor de ϵ para um valor em particular de x não está relacionado ao valor de ϵ para qualquer outro valor de x; assim, o valor de y para um valor em particular de x não está relacionado com o valor de y para qualquer outro valor de x.
4. O termo de erro ϵ é uma variável aleatória normalmente distribuída.
   Implicação: Uma vez que y é uma função linear de ϵ, y também é uma variável aleatória normalmente distribuída.

A Figura 12.6 ilustra as suposições de modelo e suas implicações; observe que, nesta interpretação gráfica, o valor de $E(y)$ se modifica de acordo com o valor específico considerado de x. Entretanto, independentemente do valor de x, a distribuição de probabilidade de ϵ e, portanto, as distribuições de probabilidade de y em qualquer ponto em particular depende de o valor real de y ser maior ou menor que $E(y)$.

Neste ponto, devemos ter em mente que também estamos fazendo uma suposição ou hipótese sobre a forma da relação entre x e y. Ou seja, supomos que a linha reta representada por $\beta_0 + \beta_1 x$ seja a base para a relação entre as variáveis. Não devemos desconsiderar o fato de que algum outro modelo, por exemplo, $y = \beta_0 + \beta_1 x^2 + \epsilon$, possa vir a ser um modelo melhor para a relação em questão.

Figura 12.6 Suposições referentes ao modelo de regressão

12.5 TESTE DE SIGNIFICÂNCIA

Em uma equação de regressão linear simples, a média, ou valor esperado, de y é uma função linear de x: $E(y) = \beta_0 + \beta_1 x$. Se o valor de b_1 for zero, $E(y) = \beta_0 + (0)x = \beta_0$. Nesse caso, o valor médio de y não depende do valor de x e, portanto, concluiríamos que x e y não estão linearmente relacionados. Alternativamente, se o valor de β_1 não for igual a zero, concluiríamos que as duas variáveis estão relacionadas. Desse modo, para testar se uma relação de regressão é significativa, devemos realizar um teste de hipóteses para determinar se o valor de β_1 é zero. Dois testes comumente são usados. Ambos requerem uma estimativa de σ^2, que é a variância de ϵ no modelo de regressão.

Estimativa de σ^2

Do modelo de regressão e de sua suposição, podemos concluir que σ^2, a variância de ϵ, também representa a variância dos valores de y nas proximidades da reta de regressão. Lembre-se de que os desvios dos valores de y nas proximidades da reta de regressão são chamados resíduos. Assim, SSE, a soma dos quadrados dos resíduos, é uma medida da variabilidade das observações reais em torno da reta de regressão estimada. O **erro médio quadrático** (MSE) fornece a estimativa de σ^2; ele é o SSE dividido por seus graus de liberdade.

Com $\hat{y}_i = b_0 + b_1 x_1$, SSE pode ser escrito como:

$$\text{SSE} = \Sigma(y_i - \hat{y}_i)^2 = \Sigma(y_i - b_0 - b_1 x_i)^2$$

Toda soma de quadrados está associada a um número que é conhecido como seus graus de liberdade. Os estatísticos demonstraram que a SSE tem $n - 2$ graus de liberdade porque dois parâmetros (β_0 e β_1) devem ser estimados para que se possa calcular SSE. Dessa forma, a média quadrática é calculada dividindo-se SSE por $n - 2$. O MSE produz um estimador sem viés de σ^2. Uma vez que o valor de MSE produz uma estimativa de σ^2, a notação σ^2 também é usada.

ERRO MÉDIO QUADRÁTICO (ESTIMATIVA DE σ^2)

$$s^2 = \text{MSE} = \frac{\text{SSE}}{n-2} \quad \textbf{(12.15)}$$

Na Seção 12.3, mostramos que, para o exemplo dos restaurantes Armand's Pizza Parlor, SSE = 1.530; portanto,

$$s^2 = \text{MSE} = \frac{1.530}{8} = 191{,}25$$

fornece uma estimativa sem viés de σ^2.

Para estimar s, extraímos a raiz quadrada de s^2. O valor resultante, s, denomina-se **erro padrão da estimativa.**

ERRO PADRÃO DA ESTIMATIVA

$$s = \sqrt{\text{MSE}} = \sqrt{\frac{\text{SSE}}{n-2}} \quad \textbf{(12.16)}$$

Para o exemplo dos restaurantes Armand's Pizza Parlors, $s = \sqrt{\text{MSE}} = \sqrt{191.25} = 13.829$. Na discussão a seguir, usaremos o erro padrão da estimativa nos testes de uma relação significativa entre x e y.

Teste t

O modelo de regressão linear simples é $y = \beta_0 + \beta_1 x + \epsilon$. Se x e y estão linearmente relacionados, devemos ter $\beta_1 \neq 0$. O propósito do teste t é verificar se podemos concluir que $\beta_1 \neq 0$. Usaremos os dados amostrais para testar a seguinte hipótese a respeito do parâmetro β_1.

$$H_0: \beta_1 = 0$$
$$H_a: \beta_1 \neq 0$$

Se H_0 for rejeitada, concluiremos que $\beta_1 \neq 0$ e que existe uma relação estatisticamente significativa entre as duas variáveis. Entretanto, se H_0 não puder ser rejeitada, teremos evidências insuficientes para concluir que existe uma relação significativa. As propriedades da distribuição amostral de b_1, o estimador por mínimos quadrados de β_1, constituem a base para o teste de hipóteses.

Em primeiro lugar, vamos considerar o que aconteceria se usássemos uma amostra aleatória diferente para o mesmo estudo de regressão. Por exemplo, suponha que a gerência do Armand's Pizza Parlors usasse os registros de vendas de uma amostra diferente de dez restaurantes. Uma análise de regressão dessa nova amostra poderia resultar em uma equação de regressão estimada similar à nossa equação de regressão estimada anterior, $\hat{y} = 60 + 5x$. Entretanto, é duvidoso que obteríamos exatamente a mesma equação (com uma interseção exatamente igual a 60 e uma inclinação exatamente igual a 5). De fato, b_0 e b_1, os estimadores por mínimos quadrados, são estatísticas amostrais que possuem suas próprias distribuições amostrais. As propriedades da distribuição amostral de b_1 são as seguintes:

DISTRIBUIÇÃO AMOSTRAL DE b_1

Valor Esperado

$$E(b_1) = \beta_1$$

Desvio Padrão

$$\sigma_{b_1} = \frac{\sigma}{\sqrt{\Sigma(x_i - \bar{x})^2}} \quad \textbf{(12.17)}$$

Forma da Distribuição

Normal

Observe que o valor esperado de b_1 é igual a β_1; então, b_1 é um estimador sem viés de β_1.

Visto que não conhecemos o valor de σ, desenvolvemos uma estimativa de σ_{b_1}, designada s_{b_1}, estimando σ com s na Equação 12.17. Desse modo, obtemos a seguinte estimativa de σ_{b_1}:

O desvio padrão de b_1 é também chamado erro padrão de b_1. Desse modo, s_{b_1} fornece uma estimativa do erro padrão de b_1.

DESVIO PADRÃO ESTIMADO DE b_1

$$s_{b_1} = \frac{s}{\sqrt{\Sigma(x_i - \bar{x})^2}} \tag{12.18}$$

Para os restaurantes Armand's Pizza Parlors, $s = 13{,}829$. Portanto, usando $\Sigma(x_i - \bar{x})^2 = 568$, como mostrado na Tabela 12.2, obtemos:

$$s_{b_1} = \frac{13{,}829}{\sqrt{568}} = 0{,}5803$$

como o desvio padrão estimado de b_1.

O teste t de uma relação significativa baseia-se no fato de a estatística de teste

$$\frac{b_1 - \beta_1}{s_{b_1}}$$

seguir uma distribuição t com $n - 2$ graus de liberdade. Se a hipótese nula for verdadeira, então $\beta_1 = 0$ e $t = b_1/s_{b_1}$.

Vamos realizar esse teste de significância para os restaurantes Armand's Pizza Parlors no nível de significância $\alpha = 0{,}01$. A estatística de teste é:

Os Apêndices 12.1 e 12.2 mostram como o Minitab e o Excel podem ser usados para calcular o valor p.

$$t = \frac{b_1}{s_{b_1}} = \frac{5}{0{,}5803} = 8{,}62$$

A tabela de distribuição t mostra que, sendo $n - 2 = 10 - 2 = 8$ graus de liberdade, $t = 3{,}355$ fornece uma área igual a 0,005 na cauda superior. Desse modo, a área na cauda superior da distribuição t correspondente à estatística de teste $t = 8{,}62$ deve ser menor que 0,005. Uma vez que esse teste é um teste bicaudal, duplicamos esse valor e concluímos que o valor p associado a $t = 8{,}62$ deve ser menor que $2(0{,}005) = 0{,}01$. O Minitab ou o Excel apresentam o valor $p = 0{,}000$. Visto que o valor p é menor que $\alpha = 0{,}01$, rejeitamos H_0 e concluímos que β_1 não é igual a zero. Essa evidência é suficiente para concluirmos que existe uma relação significativa entre a população de estudantes e as vendas trimestrais. Um resumo do teste t da significância na regressão linear simples é apresentado a seguir:

TESTE t DE SIGNIFICÂNCIA NA REGRESSÃO LINEAR SIMPLES

$$H_0: \beta_1 = 0$$
$$H_a: \beta_1 \neq 0$$

ESTATÍSTICA DE TESTE

$$t = \frac{b_1}{s_{b_1}} \tag{12.19}$$

REGRA DE REJEIÇÃO

Critério do valor p:	Rejeitar H_0 se o valor $p \leq \alpha$
Critério do valor crítico:	Rejeitar H_0 se $t \leq -t_{\alpha/2}$ ou se $t \geq t_{\alpha/2}$

em que $t_{\alpha/2}$ se baseia em uma distribuição t com $n - 2$ graus de liberdade.

Intervalo de Confiança de β_1

A forma de um intervalo de confiança de β_1 é a seguinte:

$$b_1 \pm t_{\alpha/2} s_{b_1}$$

O estimador por ponto é b_1 e a margem de erro é $t_{\alpha/2} s_{b_1}$. O coeficiente de confiança associado com esse intervalo é $1 - \alpha$, e $t_{\alpha/2}$ é o valor de t que fornece uma área igual a $\alpha/2$ na cauda superior de uma distribuição t com $n - 2$ graus de liberdade. Por exemplo, suponha que quiséssemos desenvolver uma estimação

por intervalo de confiança de β_1 para os restaurantes Armand's Pizza Parlors. Na Tabela 2 do Apêndice B, descobrimos que o valor t correspondente a $\alpha = 0{,}01$ e $n - 2 = 10 - 2 = 8$ graus de liberdade é $t_{0{,}005} =$ 3,355. A estimação por intervalo de confiança de β_1 é:

$$b_1 \pm t_{\alpha/2} s_{b_1} = 5 \pm 3{,}355(0{,}5803) = 5 \pm 1{,}95$$

ou 3,05 a 6,95.

Ao usar o teste t de significância, as hipóteses testadas foram:

$$H_0\colon \beta_1 = 0$$
$$H_a\colon \beta_1 \neq 0$$

Ao nível de significância $\alpha = 0{,}01$, podemos usar o intervalo de confiança de 99% como uma alternativa para chegar à conclusão do teste de hipóteses referente aos dados dos restaurantes Armand's Pizza Parlors. Desde que 0, o valor hipotético de β_1, não esteja incluído no intervalo de confiança (3,05 a 6,95), podemos rejeitar H_0 e concluir que existe uma relação estatística significativa entre o tamanho da população de estudantes e as vendas trimestrais. Em geral, um intervalo de confiança pode ser usado para testar qualquer hipótese bilateral a respeito de β_1. Se o valor hipotético de β_1 estiver contido no intervalo de confiança, não rejeite H_0. Caso contrário, rejeite H_0.

Teste *F*

Um teste F, baseado na distribuição F de probabilidade, também pode ser usado para testar a significância na regressão. Com somente uma variável independente, o teste F fornecerá a mesma conclusão que o teste t; ou seja, se o teste t indicar que $\beta_1 \neq 0$ e, portanto, uma relação significativa, o teste F também indicará uma relação significativa. Entretanto, com mais de uma variável independente, somente o teste F pode ser usado para testar uma relação significativa global.

A lógica subjacente ao uso do teste F para determinar se a relação de regressão é estatisticamente significativa baseia-se no desenvolvimento de duas estimativas independentes de σ^2. Explicamos como o MSE fornece uma estimativa de σ^2. Se a hipótese nula H_0: $\beta_1 = 0$ for verdadeira, a soma dos quadrados da regressão, SSR, dividida por seus graus de liberdade, produzirá outra estimativa independente de σ^2. Essa estimativa denomina-se *quadrado médio devido à regressão*, ou simplesmente *regressão pela média quadrática* (*mean square regression*), e é designada MSR. Em geral,

$$\text{MSR} = \frac{\text{SSR}}{\text{Graus de liberdade da regressão}}$$

Quanto aos modelos que consideramos neste livro, os graus de liberdade da regressão são sempre iguais ao número de variáveis independentes que há no modelo:

$$\text{MSR} = \frac{\text{SSR}}{\text{Número de variáveis independentes}} \tag{12.20}$$

Uma vez que, neste capítulo, consideramos somente modelos de regressão com uma variável independente, obtemos MSR = SSR/1 = SSR. Portanto, para os restaurantes Armand's Pizza Parlors, MSR = SSR = 14.200.

Se a hipótese nula (H_0:$\beta_1 = 0$) for verdadeira, MSR e MSE são duas estimativas independentes de σ^2 e a distribuição amostral de MSR/MSE segue uma distribuição F, sendo o grau de liberdade do numerador igual a 1 e os graus de liberdade do denominador iguais a $n - 2$. Portanto, quando $\beta_1 = 0$, o valor de MSR/MSE deve estar próximo de 1. Entretanto, se a hipótese nula for falsa ($\beta_1 \neq 0$), MSR superestimará σ^2 e o valor de MSR/MSE será inflacionado; deste modo, valores grandes de MSR/MSE levam à rejeição de H_0 e à conclusão de que a relação entre x e y é estatisticamente significativa.

Na Seção 10.4, mostramos como determinar um valor p usando a tabela de distribuição F.

Vamos concluir o teste F para o exemplo dos restaurantes Armand's Pizza Parlors. A estatística de teste é:

$$F = \frac{\text{MSR}}{\text{MSE}} = \frac{14.200}{191{,}25} = 74{,}25$$

O teste F e o teste t produzem resultados idênticos para a regressão linear simples.

A tabela de distribuição F (Tabela 4 do Apêndice B) mostra que com um grau de liberdade no numerador e $n - 2 = 10 - 2 = 8$ graus de liberdade no denominador, $F = 11{,}26$ fornece uma área igual a 0,01 na cauda superior. Desse modo, a área na cauda superior da distribuição F correspondente à estatística de teste $F = 74{,}25$ deve ser menor que 0,01. Assim, concluímos que o valor p deve ser menor que 0,01. O Minitab ou o Excel apresentam o valor $p = 0{,}000$. Uma vez que o valor p é menor que $\alpha = 0{,}01$, rejeitamos H_0 e concluímos que existe uma relação significativa entre o tamanho da população estudantil e as vendas trimestrais. Um resumo do teste F de significância na regressão linear simples é apresentado a seguir.

Se H_0 for falsa, MSE ainda assim fornecerá uma estimativa sem viés de σ^2 e MSR superestimará σ^2. Se H_0 for verdadeira, tanto MSE como MSR fornecem estimativas sem viés de σ^2; nesse caso, o valor de MSR/MSE deve ser próximo de 1.

TESTE *f* DE SIGNIFICÂNCIA NA REGRESSÃO LINEAR SIMPLES

$$H_0: \beta_1 = 0$$
$$H_a: \beta_1 \neq 0$$

ESTATÍSTICA DE TESTE

$$F = \frac{\text{MSR}}{\text{MSE}} \qquad \textbf{(12.21)}$$

REGRA DE REJEIÇÃO

Critério do valor p: Rejeitar H_0 se o valor $p \leq \alpha$
Critério do valor crítico: Rejeitar H_0 se $F \geq F_\alpha$

em que F_α se baseia em uma distribuição F com um grau de liberdade no numerador e $n - 2$ graus de liberdade no denominador.

No Capítulo 10, abordamos a análise de variância (ANOVA) e mostramos como uma **tabela ANOVA** poderia ser usada para produzir um resumo conveniente dos aspectos computacionais da análise de variância. Uma tabela ANOVA idêntica pode ser usada para resumir os resultados do teste F de significância na regressão. A Tabela 12.5 é a forma geral da tabela ANOVA para a regressão linear simples. A Tabela 12.6 é a tabela ANOVA com cálculos do teste F executados para o exemplo dos restaurantes Armand's Pizza Parlors. Regressão, Erro e Total são os rótulos das três fontes de variação, e SSR, SSE e SST aparecem como a soma de quadrados correspondente na coluna 2. Os graus de liberdade, 1 para SSR, $n - 2$ para SSE e $n - 1$ para SST, são apresentados na coluna 3. A coluna 4 contém os valores de MSR e MSE, e a coluna 5 possui o valor de $F = \text{MSR/MSE}$. Quase todas as saídas de dados de computador sobre a análise de regressão incluem um resumo do teste F de significância no formato de tabela ANOVA.

Em toda tabela de análise de variância, a soma total dos quadrados é o somatório da soma de quadrados pela regressão e a soma de quadrados dos erros; além disso, o total dos graus de liberdade é a soma dos graus de liberdade da regressão e dos graus de liberdade dos erros.

Tabela 12.5 Forma geral da tabela ANOVA para regressão linear simples

Fonte de Variação	Soma de Quadrados	Graus de Liberdade	Quadrado Médio	F
Regressão	SSR	1	$\text{MSR} = \frac{\text{SSR}}{1}$	$F = \frac{\text{MSR}}{\text{MSE}}$
Erro	SSE	$n - 2$	$\text{MSE} = \frac{\text{SSE}}{n-2}$	
Total	SST	$n - 1$		

Alguns Cuidados com a Interpretação dos Testes de Significância

A análise de regressão, a qual pode ser usada para identificar como as variáveis estão associadas entre si, não pode ser usada como evidência de uma relação de causa e efeito.

Rejeitar a hipótese nula H_0: $\beta_1 = 0$ e concluir que a relação entre x e y é significativa não nos permite concluir que existe uma relação de causa e efeito entre x e y. A conclusão de que existe uma relação de causa e efeito somente é garantida se o analista puder fornecer algum tipo de justificativa teórica de que a relação é realmente causal.

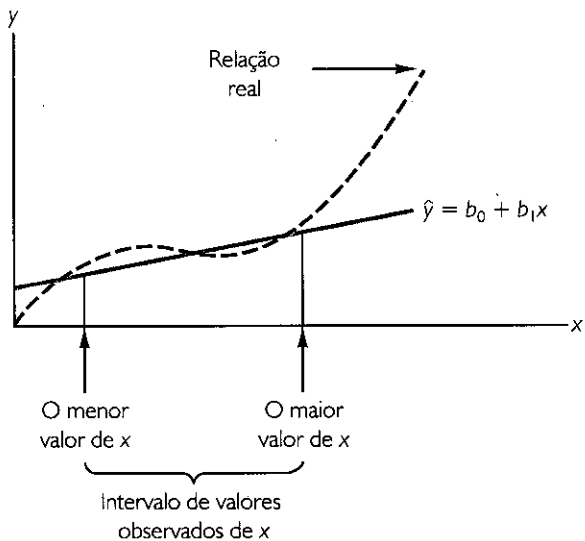
No exemplo dos restaurantes Armand's Pizza Parlors, podemos concluir que há uma relação significativa entre o tamanho da população estudantil x e as vendas trimestrais y; além disso, a equação de regressão estimada $\hat{y} = 60 + 5x$ fornece a estimativa da relação pelo método dos mínimos quadrados. Entretanto, não podemos concluir que quaisquer alterações na população estudantil x *provocam* alterações nas vendas trimestrais y simplesmente porque identificamos uma relação estatisticamente significativa. A conveniência dessa conclusão de causa e efeito reserva-se como justificativa teórica de apoio e ao bom julgamento

da parte do analista. Os gerentes dos restaurantes Armand's achavam que um aumento da população estudantil fosse uma causa provável do aumento das vendas trimestrais. Desse modo, o resultado do teste de significância lhes possibilitou concluir que havia uma relação de causa e efeito.

Além disso, simplesmente porque somos capazes de rejeitar H_0: $\beta_1 = 0$ e demonstrar a significância estatística não nos possibilita concluir que a relação entre x e y seja linear. Podemos afirmar somente que x e y estão relacionados e que uma relação linear explica a parte significativa da variabilidade em y ao longo da faixa de valores de x observados na amostra. A Figura 12.7 ilustra essa situação. O teste de significância exige a rejeição da hipótese nula H_0: $\beta_1 = 0$ e leva à conclusão de que x e y são significativamente relacionados, mas a figura mostra que a relação real entre x e y não é linear. Não obstante a aproximação linear oferecida por $\hat{y} = b_0 + b_1x$ ser boa ao longo da faixa de valores de x observados na amostra, ela torna-se fraca para valores de x fora desse intervalo.

Tabela 12.6 Tabela ANOVA para o problema dos restaurantes Armand's Pizza Parlors

Fonte de Variação	Soma de Quadrados	Graus de Liberdade	Quadrado Médio	F
Regressão	14.200	1	$\frac{14.200}{1} = 14.200$	$\frac{14.200}{191,25} = 74,25$
Erro	1.530	8	$\frac{1.530}{8} = 191,25$	
Total	15.730	9		

Figura 12.7 Exemplo de aproximação linear a uma relação não linear

Dada uma relação significativa, devemos nos sentir confiantes em usar a equação de regressão estimada para fazer previsões correspondentes a valores de x dentro do intervalo dos valores de x observados na amostra. Em relação aos restaurantes Armand's Pizza Parlors, esse intervalo corresponde a valores de x entre 2 e 26. A menos que outras razões indiquem que o modelo é válido além dessa faixa, previsões fora do intervalo da variável independente devem ser feitas com cautela. Quanto aos restaurantes Armand's Pizza Parlors, desde que a relação de regressão foi considerada significativa ao nível de 0,01, devemos nos sentir confiantes em usá-la para prever as vendas para os restaurantes em que a população estudantil correspondente está entre 2 mil e 26 mil estudantes.

NOTAS E COMENTÁRIOS

1. As suposições feitas a respeito do termo de erro (Seção 12.4) são o que possibilita os testes de significância estatística desta seção. As propriedades da distribuição amostral de b_1 e os subseqüentes testes t e F decorrem diretamente dessas suposições.
2. Não confunda significância estatística com significância prática. Com tamanhos de amostra muito grandes, resultados estatisticamente significativos podem ser obtidos para valores pequenos de b_1; nesses casos, deve-se tomar cuidado ao concluir que a relação tem significância prática.
3. Um teste de significância de uma relação linear entre x e y também pode ser executado usando-se o coeficiente de correlação da amostra r_{xy}. Com ρ_{xy} designando o coeficiente de correlação populacional, as hipóteses são as seguintes:

$$H_0: \rho_{xy} = 0$$
$$H_a: \rho_{xy} \neq 0$$

Pode-se concluir que há uma relação significativa se H_0 for rejeitada. Os detalhes desse teste são fornecidos em livros mais avançados. Entretanto, os testes t e F apresentados anteriormente nesta seção fornecem o mesmo resultado que o teste de significância usando o coeficiente de correlação. Por conseguinte, a realização de um teste de significância usando o coeficiente de correlação não é necessária se um teste t ou um teste F já tiverem sido realizados.

Exercícios

Métodos

AUTOTESTE

23. Os dados do exercício 1 são os seguintes:

x_i	1	2	3	4	5
y_i	3	7	5	11	14

a. Calcule o erro médio quadrático usando a Equação 12.15.
b. Calcule o erro padrão da estimativa usando a Equação 12.16.
c. Calcule o desvio padrão estimado de b_1 usando a Equação 12.18.
d. Use o teste t para testar as seguintes hipóteses (α = 0,05):

$$H_0: \beta_1 = 0$$
$$H_a: \beta_1 \neq 0$$

e. Use o teste F para testar as hipóteses do item (d) no nível de significância 0,05. Apresente os resultados no formato de tabela de análise de variância.

24. Os dados do exercício 2 são os seguintes:

x_i	2	3	5	1	8
y_i	25	25	20	30	16

a. Calcule o erro médio quadrático usando a Equação 12.15.
b. Calcule o erro padrão da estimativa usando a Equação 12.16.
c. Calcule o desvio padrão estimado de b_1 usando a Equação 12.18.
d. Use o teste t para testar as seguintes hipóteses (α = 0,05):

$$H_0: \beta_1 = 0$$
$$H_a: \beta_1 \neq 0$$

e. Use o teste F para testar as hipóteses do item (d) no nível de significância 0,05. Apresente os resultados no formato de tabela de análise de variância.

25. Os dados do exercício 3 são os seguintes:

x_i	2	4	5	7	8
y_i	2	3	2	6	4

a. Qual é o valor do erro padrão da estimativa?
b. Verifique se há uma relação significativa usando o teste t. Use $\alpha = 0{,}05$.
c. Use o teste F para verificar se há uma relação significativa. Use $\alpha = 0{,}05$. Qual é a sua conclusão?

Aplicações

AUTOTESTE

26. No exercício 18, os dados sobre o *grade point average* (GPA) e os salários mensais foram os seguintes:

GPA	Salário Mensal (US$)	GPA	Salário Mensal (US$)
2,6	3.300	3,2	3.500
3,4	3.600	3,5	3.900
3,6	4.000	2,9	3.600

a. O teste t indica uma relação significativa entre o GPA e o salário mensal? Qual é a sua conclusão? Use $\alpha = 0{,}05$.
b. Verifique se há uma relação significativa usando o teste F. Qual é a sua conclusão? Use $\alpha = 0{,}05$.
c. Apresente a tabela ANOVA.

27. A *Outside Magazine* testou dez diferentes modelos de mochilas *day hikers* e botas de excursão. Os dados a seguir apresentam a capacidade de resistência e o preço de cada modelo testado. A capacidade de resistência foi medida usando-se uma escala de avaliação de 1 a 5, e a avaliação 1 designa a capacidade de resistência média e uma avaliação 5 designa uma resistência excelente (*Outside Magazine Buyer's Guide*, 2001).

ARQUIVO DA INTERNET

Boots

Fábrica e Modelo	Capacidade de Resistência	Preço (US$)
Salomon Super Raid	2	120
Merrell Chameleon Prime	3	125
Teva Challenger	3	130
Vasque Fusion GTX	3	135
Boreal Maigmo	3	150
L.L. Bean GTX Super Guide	5	189
Lowa Kibo	5	190
Asolo AFX 520 GTX	4	195
Raichle Mt. Trail GTX	4	200
Scarpa Delta SL M3	5	220

a. Use esses dados para desenvolver uma equação de regressão estimada para estimar o preço de uma mochila *day hiker* e uma bota de excursão dada a avaliação da capacidade de resistência.
b. No nível de significância 0,05, determine se a capacidade de resistência e o preço estão relacionados.
c. Você se sentiria à vontade em usar a equação de regressão estimada desenvolvida no item (a) para estimar o preço de uma mochila *day hiker* ou de uma bota de excursão, dada a avaliação da capacidade de resistência?
d. Estime o preço de uma mochila *day hiker*, com a avaliação de sua capacidade de resistência sendo 4.

ARQUIVO DA INTERNET

PCs

28. Consulte o exercício 10, em que uma equação de regressão estimada relacionando a pontuação de desempenho e a avaliação global de um PC *notebook* foi desenvolvida. No nível de significância de 0,05, teste se a pontuação de desempenho e a avaliação global estão relacionadas. Apresente a tabela ANOVA. Qual é a sua conclusão?

29. Consulte o exercício 21, em que foram usados dados do volume de produção e de custo para desenvolver uma equação de regressão estimada relacionando o volume de produção e o custo de uma operação de manufatura em particular. Use $\alpha = 0{,}05$ para testar se o volume de produção está significativamente relacionado com o custo total. Apresente a tabela ANOVA. Qual é a sua conclusão?

30. Consulte o exercício 22, em que foram usados os seguintes dados para determinar se o preço de uma impressora está relacionado com a velocidade para imprimir textos simples (*PC World*, fevereiro de 2003).

ARQUIVO DA INTERNET

Printers

Marca	Tipo	Velocidade(ppm)	Preço ($)
Minolta-QMS PagePro 1250W	Pequeno Escritório	12	199
Minolta-QMS PagePro 1250W	Pequeno Escritório	12	199
Brother HL-1850	Pequeno Escritório	10	499
Lexmark E320	Pequeno Escritório	12,2	299
Minolta-QMS PagePro 1250E	Pequeno Escritório	10,3	299
HP Laserjet 1200	Pequeno Escritório	11,7	399
Xerox Phaser 4400/N	Corporativa	17,8	1.850

Marca	Tipo	Velocidade(ppm)	Preço ($)
Brother HL-2460N	Corporativa	16,1	1.000
IBM Infoprint 1120n	Corporativa	11,8	1.387
Lexmark W812	Corporativa	19,8	2.089
Oki Data B8300n	Corporativa	28,2	2.200

As evidências indicam uma relação significativa entre a velocidade de impressão e o preço? Realize o teste estatístico apropriado e declare sua conclusão. Use $\alpha = 0{,}05$.

31. Consulte o exercício 20, em que foi desenvolvida uma equação de regressão estimada relacionando a renda familiar típica e o preço típico de uma moradia. Teste se a renda familiar típica de uma cidade e o preço típico de uma moradia estão relacionados ao nível de significância 0,01.

12.6 USANDO A EQUAÇÃO DE REGRESSÃO ESTIMADA PARA ESTIMAÇÃO E PREVISÃO

Quando usamos o modelo de regressão linear simples estamos fazendo uma suposição sobre a relação entre x e y. Então, usamos o método dos mínimos quadrados para obter a equação de regressão linear simples estimada. Se existir uma relação significativa entre x e y, e se o coeficiente de determinação mostrar que o ajuste é bom, a equação de regressão estimada será útil para estimação e previsão.

Estimação por Ponto

No exemplo dos restaurantes Armand's Pizza Parlors, a equação de regressão estimada $\hat{y} = 60 + 5x$ fornece uma estimativa da relação entre o tamanho da população estudantil x e as vendas trimestrais y. Podemos usar a equação de regressão estimada para desenvolver uma estimação por ponto do valor médio de y para um valor em particular de x ou para prever um valor individual de y correspondente a determinado valor de x. Por exemplo, suponha que os gerentes dos restaurantes Armand's queiram uma estimação por ponto da média de vendas trimestrais de todos os restaurantes localizados nas proximidades de *campi* universitários que possuam 10 mil estudantes. Usando a equação de regressão estimada $\hat{y} = 60 + 5x$, notamos que para $x = 10$ (ou 10 mil estudantes), $\hat{y} = 60 + 5(10) = 110$. Desse modo, uma estimação por ponto da média das vendas trimestrais para todos os restaurantes localizados próximo a *campi* universitários com 10 mil estudantes é US$ 110 mil.

Suponha agora que os gerentes do Armand's queiram prever as vendas relativas a um determinado restaurante localizado próximo ao Talbot College, uma escola com 10 mil estudantes. Nesse caso, não estamos interessados no valor médio de todos os restaurantes localizados perto de *campi* universitários com 10 mil estudantes; estamos apenas interessados em prever as vendas trimestrais de um determinado restaurante. Ocorre que a estimação por ponto de um valor individual de y é a mesma estimação por ponto referente ao valor médio de y. Portanto, preveríamos vendas trimestrais de $\hat{y} = 60 + 5(10) = 110$, ou US$ 110 mil para esse restaurante em particular.

Os intervalos de confiança e os intervalos de previsão mostram a precisão dos resultados da regressão. Intervalos mais estreitos produzem um grau mais elevado de precisão.

Estimação por Intervalo

A estimação por ponto não fornece nenhuma informação sobre a precisão associada a uma estimativa. Para tanto, precisamos desenvolver estimações por intervalo muito similares às dos Capítulos 8, 10 e 11. O primeiro tipo de estimação por intervalo, um **intervalo de confiança**, é uma estimação por intervalo do *valor médio de y* para determinado valor de x. O segundo tipo de estimação por intervalo, um **intervalo de previsão**, é usado quando queremos uma estimação por intervalo de um *valor individual de y* para determinado valor de x. A estimação por ponto do valor médio de y é similar à estimação por ponto de um valor individual de y. Porém, as estimações por intervalo que obtemos para os dois casos são diferentes. A margem de erro é maior para um intervalo de previsão.

Intervalo de Confiança do Valor Médio de y

A equação de regressão estimada fornece uma estimação por ponto do valor médio de y para determinado valor de x. Para desenvolver o intervalo de confiança, usaremos a seguinte notação:

x_p = o valor em particular ou determinado da variável independente x
y_p = o valor da variável dependente y correspondente ao x_p dado
$E(y_p)$ = o valor médio ou esperado da variável dependente y correspondente ao x_p dado
$\hat{y}_p = b_0 + b_1x_p$ = a estimação por ponto de $E(y_p)$ quando $x = x_p$

Usando essa notação para estimar a média das vendas de todos os restaurantes Armand's localizados próximo a um *campus* universitário com 10 mil estudantes, temos $x_p = 10$, e $E(y_p)$ designa o valor médio desconhecido das vendas correspondentes a todos os restaurantes onde $x_p = 10$. A estimação por ponto de $E(y_p)$ é fornecida por $\hat{y}_p = 60 + 5(10) = 110$.

Em geral, não podemos esperar que $\hat{y}_p$ seja exatamente igual a $E(y_p)$. Se quisermos fazer uma inferência sobre quão próximo $\hat{y}_p$ está do verdadeiro valor médio $E(y_p)$, teremos de estimar a variância de $\hat{y}_p$. A fórmula para estimar a variância de $\hat{y}_p$, dado x_p, designada por $s^2_{\hat{y}_p}$ é:

$$s^2_{\hat{y}_p} = s^2\left[\frac{1}{n} + \frac{(x_p - \bar{x})^2}{\Sigma(x_i - \bar{x})^2}\right] \qquad (12.22)$$

A estimativa do desvio padrão de $\hat{y}_p$ é dada pela raiz quadrada da Equação 12.22.

$$s_{\hat{y}_p} = s\sqrt{\frac{1}{n} + \frac{(x_p - \bar{x})^2}{\Sigma(x_i - \bar{x})^2}} \qquad (12.23)$$

Os resultados computacionais referentes ao Armand's Pizza Parlors da Seção 12.5 forneceram $s = 13{,}829$. Com $x_p = 10$, $\bar{x} = 14$, e $\Sigma(x_i - \bar{x})^2 = 568$, podemos usar a Equação 12.23 para obter:

$$s_{\hat{y}_p} = 13{,}829\sqrt{\frac{1}{10} + \frac{(10 - 14)^2}{568}}$$
$$= 13{,}829\sqrt{0{,}1282} = 4{,}95$$

A expressão geral para um intervalo de confiança é o seguinte:

A margem de erro associada a esta estimação de intervalo é $t_{\alpha/2}s_{\hat{y}_p}$.

INTERVALO DE CONFIANÇA DE $E(y_p)$

$$\hat{y}_p \pm t_{\alpha/2}s_{\hat{y}_p} \qquad (12.24)$$

em que o coeficiente de confiança é $1 - \alpha$ e $t_{\alpha/2}$ baseia-se em uma distribuição t com $n - 2$ graus de liberdade.

Usando a Equação 12.24 para desenvolver um intervalo de confiança de 95% da média de vendas trimestrais de todos os restaurantes Armand's localizados próximos a *campi* universitários com 10 mil estudantes, precisamos do valor t para $\alpha/2 = 0{,}025$ e $n = 2 = 10 - 2 = 8$ graus de liberdade. Usando a Tabela 2 do Apêndice B, obtemos $t_{0{,}025} = 2{,}306$. Desse modo, com $\hat{y}_p = 110$ e uma margem de erro de $t_{\alpha/2}s_{\hat{y}_p} = 2{,}306(4{,}95) = 11{,}415$, a estimação por intervalo de confiança de 95% é:

$$110 \pm 11{,}415$$

Em termos de dólares, o intervalo de confiança de 95% da média das vendas trimestrais de todos os restaurantes próximos a *campi* universitários com 10 mil estudantes é US$ 110 mil ± US$ 11.415. Portanto, o intervalo de confiança de 95% correspondente à média das vendas trimestrais quando a população estudantil é de 10 mil alunos varia de US$ 98.585 a US$ 121.415.

Observe que o desvio padrão estimado de $\hat{y}_p$ dado pela Equação 12.23 é menor quando $x_p = \bar{x}$ e a quantidade $x_p = \bar{x} = 0$. Nesse caso, o desvio padrão estimado de $\hat{y}_p$ torna-se:

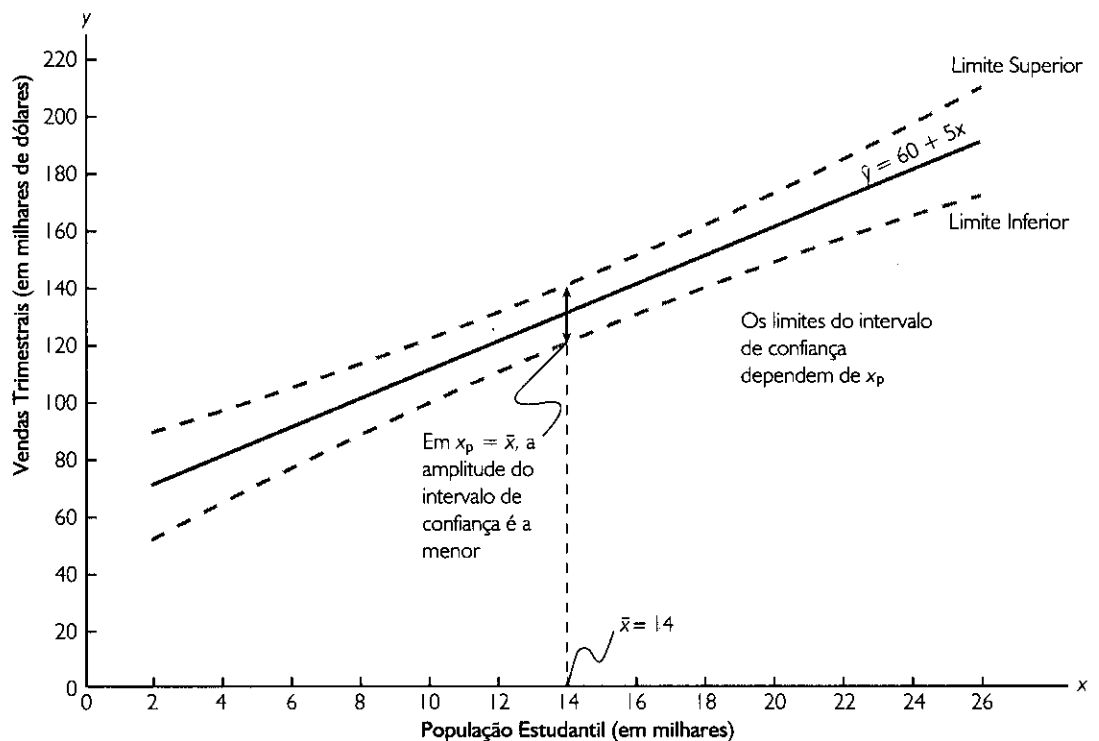
$$s_{\hat{y}_p} = s\sqrt{\frac{1}{n} + \frac{(\bar{x} - \bar{x})^2}{\Sigma(x_i - \bar{x})^2}} = s\sqrt{\frac{1}{n}}$$

Esse resultado implica que podemos fazer a melhor ou a mais precisa estimativa do valor médio de y quando quer que $x_p = \bar{x}$. Realmente, quanto mais distante x_p estiver de $\bar{x}$, maior $x_p - \bar{x}$ se torna. Em con-

seqüência, os intervalos de confiança do valor médio de y se tornarão mais amplos quanto mais x_p se desvia de $\bar{x}$. Esse padrão é mostrado graficamente na Figura 12.8.

Intervalo de Previsão para um Valor Individual de y

Suponha que, em vez de estimar o valor médio das vendas correspondentes a todos os restaurantes Armand's, localizados nas proximidades de *campi* universitários com 10 mil estudantes, queiramos estimar as vendas correspondentes a um determinado restaurante, localizado próximo ao Talbot College, uma escola com 10 mil estudantes. Conforme observamos anteriormente, a estimação por ponto de $\hat{y}_p$, o valor de y correspondente ao x_p dado, é fornecida pela equação de regressão estimada $\hat{y}_p = b_0 + b_1x_p$. Quanto ao restaurante do Talbot College, temos $x_p = 10$, e as correspondentes vendas trimestrais previstas são $\hat{y}_p = 60 + 5(10) = 110$, ou US\$ 110 mil. Note que esse valor é idêntico à estimação por ponto da média das vendas correspondentes a todos os restaurantes localizados nas proximidades de *campi* universitários com 10 mil estudantes.

Figura 12.8 Intervalos de confiança da média de vendas y a dados valores da população estudantil x

Para desenvolver um intervalo de previsão, devemos primeiramente determinar a variância associada ao uso de $\hat{y}_p$ como estimativa de um valor individual de y quando $x = x_p$. Essa variância é composta da soma dos dois componentes seguintes:

1. A variância dos valores individuais de y nas proximidades da média $E(y_p)$, uma estimativa da qual é dada por s^2.
2. A variância associada ao uso de $\hat{y}_p$ para estimar $E(y_p)$, uma estimativa da qual é dada por $s^2_{\hat{y}_p}$.

A fórmula para estimar a variância de um valor individual de y_p, designada por s^2_{ind} é:

$$s^2_{\text{ind}} = s^2 + s^2_{\hat{y}_p}$$

$$= s^2 + s^2\left[\frac{1}{n} + \frac{(x_p - \bar{x})^2}{\Sigma(x_i - \bar{x})^2}\right]$$

$$= s^2\left[1 + \frac{1}{n} + \frac{(x_p - \bar{x})^2}{\Sigma(x_i - \bar{x})^2}\right] \tag{12.25}$$

Portanto, uma estimativa do desvio padrão de um valor individual de y_p é dada por:

$$s_{ind} = s\sqrt{1 + \frac{1}{n} + \frac{(x_p - \bar{x})^2}{\Sigma(x_i - \bar{x})^2}} \qquad (12.26)$$

Quanto aos restaurantes Armand's Pizza Parlors, o desvio padrão estimado correspondente à previsão de vendas para um restaurante específico, localizado perto de um *campus* universitário com 10 mil estudantes, é calculado da seguinte maneira:

$$\begin{aligned} s_{ind} &= 13{,}829\sqrt{1 + \frac{1}{10} + \frac{(10 - 14)^2}{568}} \\ &= 13{,}829\sqrt{1{,}1282} \\ &= 14{,}69 \end{aligned}$$

A expressão geral de um intervalo de previsão é a seguinte:

INTERVALO DE PREVISÃO DE y_p

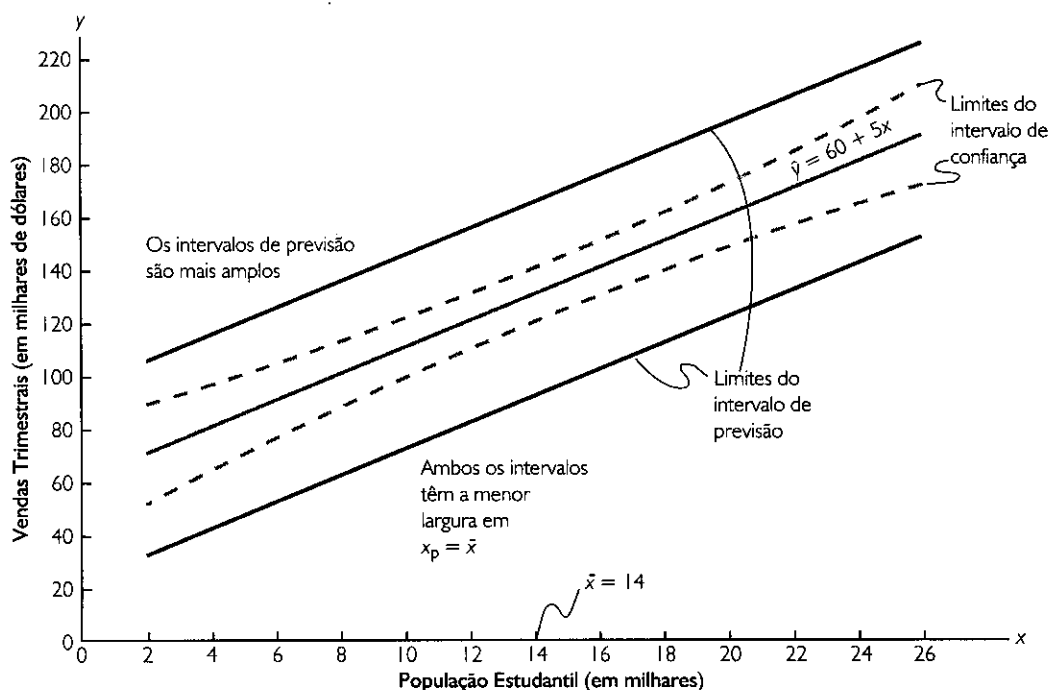
$$\hat{y}_p \pm t_{\alpha/2} s_{ind} \qquad (12.27)$$

em que o coeficiente de confiança é $1 - \alpha$ e $t_{\alpha/2}$ baseia-se em uma distribuição t com $n - 2$ graus de liberdade.

O intervalo de previsão de 95% relativo às vendas trimestrais no restaurante do Talbot College pode ser encontrado usando-se $t_{0,025} = 2{,}306$ e $s_{ind} = 14{,}69$. Desse modo, com $\hat{y}_p = 110$ e uma margem de erro igual a $t_{\alpha/2} s_{ind} = 2{,}306(14{,}69) = 33{,}875$, o intervalo de previsão de 95% é:

$$110 \pm 33{,}875$$

A margem de erro associada a essa estimação por intervalo é $t_{\alpha/2} s_{ind}$.

Figura 12.9 Intervalos de confiança e de previsão de vendas y a dados valores da população estudantil x

Em termos de dólares, esse intervalo de previsão é US\$ 110 mil ± US\$ 33.875 ou US\$ 76.125 a US\$ 143.875. Observe que o intervalo de previsão para um restaurante em particular localizado próximo a um *campus* com 10 mil estudantes é mais amplo que o intervalo de confiança para a média de vendas de todos os restaurantes localizados próximo o *campi* com 10 mil estudantes. A diferença reflete o fato de sermos capazes de estimar o valor médio de y mais precisamente do que um valor individual de y.

Tanto as estimações por intervalo de confiança como as estimações por intervalo de previsão são mais precisas quando o valor da variável independente é $x_p = \bar{x}$. As formas gerais dos intervalos de confiança e os intervalos de previsão mais amplos são mostrados juntos na Figura 12.9.

Exercícios

Métodos

AUTOTESTE

32. Os dados do exercício 1 são os seguintes:

x_i	1	2	3	4	5
y_i	3	7	5	11	14

a. Use a Equação 12.23 para estimar o desvio padrão de $\hat{y}_p$ quando $x = 4$.
b. Use a Equação 12.24 para desenvolver um intervalo de confiança de 95% para o valor esperado de y quando $x = 4$.
c. Use a Equação 12.26 para estimar o desvio padrão de um valor individual de y quando $x = 4$.
d. Use a Equação 12.27 para desenvolver um intervalo de previsão de 95% para y quando $x = 4$.

33. Os dados do exercício 2 são os seguintes:

x_i	2	3	5	1	8
y_i	25	25	20	30	16

a. Estime o desvio padrão de $\hat{y}_p$ quando $x = 3$.
b. Desenvolva um intervalo de confiança de 95% para o valor esperado de y quando $x = 3$.
c. Estime o desvio padrão de um valor individual de y quando $x = 3$.
d. Estipule um intervalo de previsão de 95% para y quando $x = 3$.

34. Os dados do exercício 3 são os seguintes:

x_i	2	4	5	7	8
y_i	2	3	2	6	4

Desenvolva os intervalos de confiança e de previsão de 95% quando $x = 3$. Explique por que esses dois intervalos são diferentes.

Aplicações

AUTOTESTE

35. No exercício 18, os dados sobre o *grade average point* (GPA) x e o salário mensal y produziram a equação de regressão estimada $\hat{y} = 1.790{,}5 + 581{,}1x$.

a. Desenvolva um intervalo de confiança de 95% referente ao salário inicial médio de todos os estudantes com uma pontuação GPA igual a 3,0.
b. Estabeleça um intervalo de previsão do salário inicial de Joe Heller, um estudante com um GPA igual a 3,0.

ARQUIVO DA INTERNET
PCs

36. No exercício 10, dados sobre a avaliação do desempenho (x) e a classificação geral (y) de computadores *notebook* forneceram a equação de regressão estimada $\hat{y} = 51.819 + 0{,}1452x$ (*PC World*, fevereiro de 2000).

a. Realize uma estimação por ponto da classificação geral de um PC com uma avaliação de desempenho igual a 200.
b. Desenvolva um intervalo de confiança de 95% para a média de avaliação global de todos os PCs que obtiveram uma avaliação de desempenho igual a 200.
c. Suponha que um novo PC desenvolvido pela Dell tenha uma avaliação de desempenho igual a 200. Desenvolva um intervalo de previsão de 95% para a avaliação global desse novo PC.
d. Discuta as diferenças em suas respostas aos itens (b) e (c).

37. No exercício 13, foram fornecidos dados sobre a renda bruta ajustada x e o valor das deduções detalhadas feitas pelos contribuintes. Os dados foram expressos em milhares de dólares. Com a equação de regressão estimada $\hat{y} = 4{,}68 + 0{,}16x$, a estimação por ponto de um nível razoável de deduções detalhadas para um contribuinte que tem uma renda bruta ajustada de US$ 52.500 é US$ 13.080.
   a. Desenvolva um intervalo de confiança de 95% do valor médio de deduções detalhadas para todos os contribuintes que tenham uma renda bruta ajustada de US$ 52.500.
   b. Faça uma estimação por intervalo de previsão de 95% do valor das deduções totais detalhadas para um contribuinte em particular que tem uma renda bruta ajustada de US$ 52.500.
   c. Se o contribuinte em particular citado no item (b) reivindicasse deduções totais detalhadas de US$ 20.400, seria justificável que o fiscal da Receita Federal requeresse uma auditoria?
   d. Use suas respostas do item (b) para dar ao fiscal da Receita Federal uma diretriz quanto ao valor de deduções totais detalhadas que um contribuinte com uma renda bruta ajustada de US$ 52.500 deveria reivindicar antes que uma auditoria seja recomendada.

38. Consulte o exercício 21, no qual foram usados dados sobre o volume de produção x e o custo total y para uma operação de manufatura em particular, para desenvolver uma equação de regressão estimada $\hat{y} = 1.246{,}67 + 7{,}6x$.
   a. O programa de produção da empresa mostra que 500 unidades devem ser produzidas no próximo mês. Qual é a estimação por ponto do custo total para o próximo mês?
   b. Desenvolva um intervalo de previsão de 99% do custo total para o próximo mês.
   c. Se um relatório contábil de custos no fim do próximo mês mostrar que o custo de produção real durante o mês foi de US$ 6 mil, os gerentes devem preocupar-se em se sujeitar a esse elevado custo total durante o mês? Discuta.

39. Quase todos os sistemas de *light-rail*[6] nos Estados Unidos usam carros elétricos que circulam em trilhos construídos no nível das ruas. A Federal Transit Administration afirma que viajar pelo *light-rail* é um dos meios mais seguros, com um índice de acidentes de 0,99 por milhão de milhas-passageiro em comparação com 2,29 para os ônibus. Os dados a seguir apresentam os quilômetros de trilhos construídos e o número estimado de passageiros nos dias úteis nos seis sistemas de *light-rail* (*USA Today*, 7 de janeiro de 2003).

Cidade	Quilômetros de Trilhos	Número Estimado de Passageiros (em milhares)
Cleveland	24,14	15
Denver	27,36	35
Portland	61,15	81
Sacramento	33,79	31
San Diego	75,64	75
San Jose	49,89	30
St. Louis	54,71	42

   a. Use os dados para desenvolver uma equação de regressão estimada que possa ser usada para prever o número de passageiros, dados os quilômetros de trilhos construídos.
   b. A equação de regressão estimada proporciona um bom ajuste? Explique.
   c. Desenvolva um intervalo de confiança de 95% para o número médio de passageiros em dias úteis para todos os sistemas de *light-rail* com 48,28 quilômetros de trilhos.
   d. Suponha que a cidade de Charlotte esteja considerando a construção de um sistema de *light-rail* com 48,28 km de trilhos. Desenvolva um intervalo de previsão de 95% correspondente ao número de passageiros em dias úteis para o sistema de Charlotte. Você acha que o intervalo de previsão que desenvolveu teria valor para os planejadores de Charlotte ao antecipar o número de viajantes nos dias úteis em seu novo sistema de *light-rail*? Explique.

[6] NT: *Light-rail* – Meio de transporte ferroviário urbano que usa bondes ou trens de pequeno porte.

12.7 SOLUÇÃO COMPUTADORIZADA

Realizar os cálculos da análise de regressão sem a ajuda de um computador pode consumir muito tempo. Nesta seção, discutiremos como o volume de cálculos pode ser minimizado usando-se um software de computador como o Minitab.

Inserimos os dados correspondentes à população estudantil e às vendas trimestrais dos restaurantes Armand's em uma planilha do Minitab. A variável independente foi intitulada Pop e a variável independente foi chamada Sales para ajudar na interpretação do impresso de computador. Usando o Minitab, obtivemos o impresso mostrado na Figura 12.10* relativo aos restaurantes Armand's Pizza Parlors. A interpretação desse impresso é a seguinte:

1. O Minitab imprime a equação de regressão estimada como Sales (Vendas) = 60,0 + 5,00 Pop.
2. É impressa uma tabela que exibe os valores do coeficiente b_0 e b_1, o desvio padrão de cada coeficiente, o valor t obtido ao dividir-se cada coeficiente por seu desvio padrão, e o valor p associado a cada teste t. Uma vez que o valor p correspondente a $b_1 = 5{,}0000$ é zero (para três casas decimais), os resultados amostrais indicam que a hipótese nula (H_0:$\beta_1 = 0$) deve ser rejeitada. Alternativamente, poderíamos comparar 8,62 (localizado na coluna razão t) com o valor crítico apropriado. Esse procedimento para o teste t foi descrito na Seção 12.5.

Figura 12.10 Saída de dados do Minitab para o problema dos restaurantes Armand's Pizza Parlors

```
The regression equation is
Sales = 60.0 + 5.00 Pop                    <---- Equação de regressão estimada

Predictor     Coef    SE Coef      T       p
Constant     60.000    9.226     6.50   0.000
Pop          5.0000   0.5803     8.62   0.000

S = 13.83     R-sq = 90.3%     R-sq(adj) = 89.1%

Analysis of Variance

SOURCE          DF      SS      MS       F       p
Regression       1   14200   14200   74.25   0.000
Residual Error   8    1530     191                  <---- Tabela ANOVA
Total            9   15730

Predicted Values for New Observations

New
Obs     Fit    SE Fit        95% C.I.          95% P.I.        <---- Estimativas de intervalo
  1  110.00      4.95   (98.58, 121.42)   (76.12, 143.88)
```

3. O Minitab imprime o erro padrão da estimativa, $s = 13{,}83$, bem como a informação sobre a eficiência de ajuste. Observe que "R-sq = 90,3%" é o coeficiente de determinação expresso na forma de porcentagem.
4. A tabela ANOVA é impressa abaixo do cabeçalho Analysis of Variance. O Minitab usa o rótulo Residual Error (Erro Residual) para a fonte de erro de variação. Observe que DF é uma abreviação de *degrees of freedom* (graus de liberdade) e que MSR é dada como 14.200 e MSE como 191. A razão desses dois valores fornece o valor F 74,25 e o valor p correspondente 0,000. Uma vez que o valor p é zero (para três casas decimais), a relação entre Sales e Pop é considerada estatisticamente significativa.

* As etapas do Minitab necessárias para gerar a saída de dados (*output*) são apresentadas no Apêndice 12.1.

5. A estimação por intervalo de confiança de 95% das vendas esperadas e a estimação por intervalo de previsão de 95% das vendas correspondentes a um restaurante individual, localizado próximo a um *campus* com 10 mil estudantes, estão impressas abaixo da tabela ANOVA.O intervalo de confiança é (98,58, 121,42) e o intervalo de previsão é (76,12, 143,88), conforme mostramos na Seção 12.6.

Exercícios

Aplicações

AUTOTESTE

40. A divisão comercial de uma empresa imobiliária está realizando uma análise de regressão da relação entre x, que são os aluguéis anuais brutos (em milhares de dólares), e y, o preço de venda (em milhares de dólares) de prédios de apartamento. Os dados foram coletados de diversas propriedades vendidas recentemente e a seguinte saída de computador foi obtida:

```
The regression equation is
Y = 20.0 + 7.21 X

Predictor      Coef     SE Coef       T
Constant     20.000      3.2213    6.21
X             7.210      1.3626    5.29

Analysis of Variance

SOURCE            DF         SS
Regression         1    41587.3
Residual Error     7
Total              8    51984.1
```

a. Quantos prédios de apartamento constavam na amostra?
b. Escreva a equação de regressão estimada.
c. Qual é o valor de s_{b_1}?
d. Use a estatística F para testar a significância da relação ao nível de significância 0,05.
e. Estime o preço de venda de um prédio de apartamento com aluguéis anuais brutos de US$ 50 mil.

41. Apresentamos a seguir uma parte da saída de dados de computador de uma análise de regressão que relaciona y = as despesas de manutenção (em dólares por mês) com x = o uso (em horas por semana) de uma marca em particular de terminal de computador.

```
The regression equation is
Y = 6.1092 + .8951 X

Predictor      Coef     SE Coef
Constant     6.1092      0.9361
X            0.8951      0.1490

Analysis of Variance

SOURCE            DF         SS         MS
Regression         1    1575.76    1575.76
Residual Error     8     349.14      43.64
Total              9    1924.90
```

a. Escreva a equação de regressão estimada.
b. Use o teste t para determinar se as despesas mensais de manutenção se relacionam com o uso ao nível de significância 0,05.

c. Use a equação de regressão estimada para prever as despesas mensais de manutenção de qualquer terminal que seja usado 25 horas por semana.

42. Um modelo de regressão relacionando x, que é o número de vendedores em uma filial, com y, as vendas anuais nessa filial (em milhares de dólares), forneceu a seguinte saída de computador de uma análise de regressão dos dados:

```
The regression equation is
Y = 80.0 + 50.00 X

Predictor      Coef     SE Coef      T
Constant       80.0      11.333    7.06
X              50.0       5.482    9.12

Analysis of Variance

SOURCE           DF       SS         MS
Regression        1     6828.6     6828.6
Residual Error   28     2298.8       82.1
Total            29     9127.4
```

a. Escreva a equação de regressão estimada.
b. Quantos escritórios filiais estavam envolvidos no estudo?
c. Calcule a estatística F e teste a significância da relação ao nível de significância 0,05.
d. Preveja as vendas anuais no escritório filial de Memphis. Essa filial emprega 12 vendedores.

43. Especialistas da área da Saúde recomendam que os corredores bebam 120 ml de água a cada 15 minutos de corrida. Não obstante as garrafas manuais funcionarem bem para muitos tipos de corrida, as corridas *cross-country*, que são feitas durante o dia inteiro, requerem sistemas de hidratação adaptados à cintura do atleta ou fixados aos ombros. Além disso, para carregar mais água, os sistemas adaptados à cintura ou fixados aos ombros oferecem mais espaço de armazenagem para alimentos e roupas extras. À medida que a capacidade aumenta, entretanto, o peso e o custo desses sistemas de maior capacidade também aumentam. Os dados seguintes apresentam o peso (em gramas) e o preço de 26 sistemas de hidratação que se adaptam à cintura ou que se fixam aos ombros do atleta (*Trail Runner Gear Guide*, 2003).

ARQUIVO DA INTERNET
Hydration1

Modelo	Peso (gramas)	Preço (US$)
Fastdraw	85	10
Fastdraw Plus	113	12
Fitness	142	12
Access	198	20
Access Plus	227	25
Solo	255	25
Serenade	225	35
Solitaire	312	35
Gemini	595	45
Shadow	425	40
SipStream	510	60
Express	255	30
Lightning	340	40
Elite	397	60
Extender	454	65
Stinger	454	65
GelFlask Belt	85	20
GelDraw	28	7
GelFlask Clip-on Holster	56	10
GelFlask Holster SS	28	10
Strider (W)	227	30

Modelo	Peso (gramas)	Preço (US$)
Walkabout (W)	397	40
Solitude I.C.E.	255	35
Getaway I.C.E.	539	55
Profile I.C.E.	397	50
Traverse I.C.E.	367	60

a. Use os dados para desenvolver uma equação de regressão estimada que possa ser usada para prever o preço de um sistema de hidratação dado o seu peso.
b. Teste a significância da relação ao nível de significância 0,05.
c. A equação de regressão estimada proporcionou um bom ajuste? Explique.
d. Suponha que a equação de regressão estimada desenvolvida no item (a) também se aplique a sistemas de hidratação produzidos por outras empresas. Desenvolva uma estimação por intervalo de confiança de 95% do preço de todos os sistemas de hidratação que pesam 283 g.
e. Suponha que a equação de regressão estimada desenvolvida no item (a) também se aplique a sistemas de hidratação produzidos por outras empresas. Desenvolva uma estimação por intervalo de previsão do preço do sistema Back Draft produzido pela Eastern Mountain Sports. O sistema Back Draft pesa 283 g.

44. A Cushman & Wakefield, Inc. coleta dados que mostram o índice de vagas em prédios de escritório e os preços de aluguel de estabelecimentos comerciais nos Estados Unidos. Os dados a seguir apresentam os índices gerais de vagas (%) e os preços médios de aluguel (por pé quadrado[7]) no centro comercial de 18 mercados selecionados.

ARQUIVO DA INTERNET
OffRates

Mercado	Índice de Vagas (%)	Preço Médio (US$)
Atlanta	21,9	18,54
Boston	6,0	33,70
Hartford	22,8	19,67
Baltimore	18,1	21,01
Washington	12,7	35,09
Philadelphia	14,5	19,41
Miami	20,0	25,28
Tampa	19,2	17,02
Chicago	16,0	24,04
San Francisco	6,6	31,42
Phoenix	15,9	18,74
San Jose	9,2	26,76
West Palm Beach	19,7	27,72
Detroit	20,0	18,20
Brooklyn	8,3	25,00
Downtown, NY	17,1	29,78
Midtown, NY	10,8	37,03
Midtown South, NY	11,1	28,64

a. Desenvolva um diagrama de dispersão desses dados. Trace o índice de vagas no eixo horizontal.
b. Parece haver alguma relação entre os índices de vagas e os preços de aluguel?
c. Desenvolva a equação de regressão estimada que possa ser usada para prever a média dos preços de aluguel, dado o índice global de vagas.
d. Teste a significância da relação ao nível de significância 0,05.
e. A equação de regressão estimada proporcionou um bom ajuste? Explique.
f. Preveja o preço de aluguel esperado para mercados com um índice de vagas de 25% no centro comercial da cidade.
g. O índice global de vagas no centro comercial de Ft. Lauderdale é 11,3%. Preveja o preço de aluguel esperado para Ft. Lauderdale.

[7] NT: 1 pé quadrado – 929,03 cm quadrados.

12.8 ANÁLISE RESIDUAL: VALIDANDO SUPOSIÇÕES DO MODELO

A análise residual é a principal ferramenta para determinar se o modelo de regressão proposto é apropriado.

Conforme observamos anteriormente, o *resíduo* da observação i é a diferença entre o valor observado da variável dependente (y_i) e o valor estimado da variável dependente ($\hat{y}_i$).

RESÍDUO DA OBSERVAÇÃO i

$$y_i - \hat{y}_i \tag{12.28}$$

em que

y_i = o valor observado da variável dependente
$\hat{y}_i$ = o valor estimado da variável dependente

Em outras palavras, o i-ésimo resíduo é o erro resultante de se usar a equação de regressão estimada para prever o valor da variável dependente. Os resíduos, no exemplo dos restaurantes Armand's Pizza Parlors, estão calculados na Tabela 12.7. Os valores observados da variável dependente estão na segunda coluna e os valores estimados da variável dependente, obtidos usando-se a equação de regressão estimada $\hat{y} = 60 + 5x$, estão na terceira coluna. Uma análise dos resíduos correspondentes na quarta coluna ajudará a determinar se as suposições feitas a respeito do modelo de regressão são apropriados.

Vamos rever agora as suposições de regressão do exemplo dos restaurantes Armand's Pizza Parlors. Presumimos um modelo de regressão linear simples:

$$y = \beta_0 + \beta_1 x + \epsilon \tag{12.29}$$

Esse modelo indica que as vendas trimestrais presumidas (y) são uma função linear do tamanho da população estudantil (x) mais um termo de erro ϵ. Na Seção 12.4, fizemos as seguintes suposições sobre o termo de erro ϵ.

1. $E(\epsilon) = 0$.
2. A variância de ϵ, designada por σ^2, é idêntica para todas as variáveis de x.
3. Os valores de ϵ são independentes.
4. O termo de erro ϵ tem uma distribuição normal.

Essas suposições constituem a base teórica para o teste t e para o teste F usados para determinar se a relação entre x e y é significativa, e para as estimações por intervalo de confiança e de previsão apresentadas na Seção 12.6. Se as suposições sobre o termo de erro ϵ parecerem questionáveis, os testes de hipótese sobre a significância da relação de regressão e os resultados da estimação por intervalo podem não ser válidos.

Os resíduos fornecem a melhor informação sobre ϵ; portanto, uma análise dos resíduos é um passo importante para determinar se as suposições referentes a ϵ são apropriadas.

Grande parte da análise residual baseia-se em um exame das plotagens gráficas. Nesta seção, discutiremos as seguintes plotagens residuais:

Tabela 12.7 Resíduos referentes ao exemplo dos restaurantes Armand's Pizza Parlors

População Estudantil x_i	Vendas y_i	Estimativa de Vendas $y_i = 60 + 5x_i$	Resíduos $y_i - y_i$
2	58	70	212
6	105	90	15
8	88	100	212
8	118	100	18
12	117	120	23
16	137	140	23
20	157	160	23
20	169	160	9
22	149	170	221
26	202	190	12

1. Uma plotagem dos resíduos em relação aos valores da variável independente x.
2. Uma plotagem dos resíduos em relação aos valores previstos da variável independente $\hat{y}$.

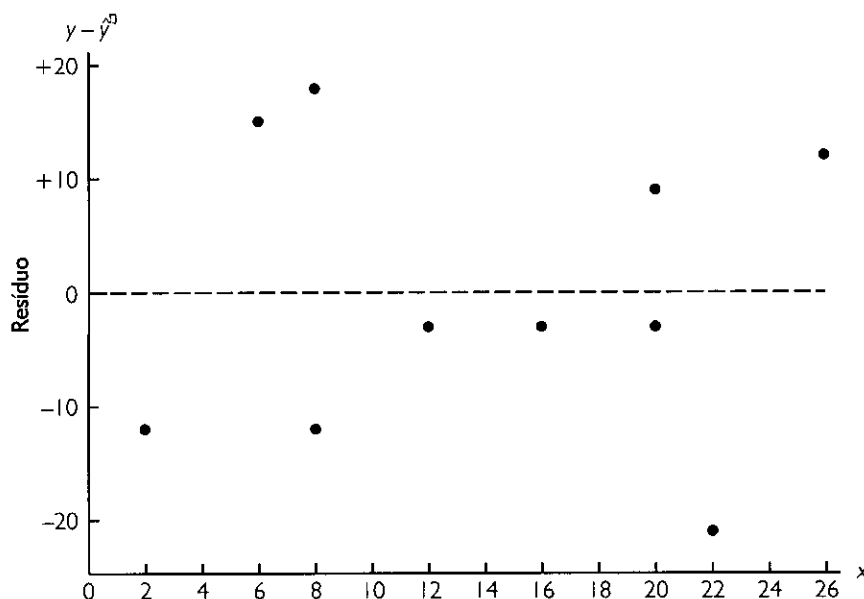
Plotagem Residual em Relação a *x*

Uma **plotagem residual** em relação à variável independente x é um gráfico no qual os valores da variável independente são representados pelo eixo horizontal e os valores residuais correspondentes são representados pelo eixo vertical. Para cada resíduo é marcado um ponto no gráfico. A primeira coordenada de cada ponto é dada pelo valor de x_i e a segunda coordenada é dada pelo correspondente valor do resíduo $y_i - \hat{y}_i$. Para uma plotagem residual em relação a x com os dados dos restaurantes Armand's Pizza Parlors da Tabela 12.7, as coordenadas do primeiro ponto são (2, –12), correspondentes a $x_1 = 2$ e $y_i - \hat{y}_i = -12$; as coordenadas do segundo ponto são (6, 15), correspondentes a $x_2 = 6$ e $y_2 - \hat{y}_2 = 15$ e assim por diante. A Figura 12.11 apresenta a plotagem residual resultante.

Antes de interpretar os resultados dessa plotagem residual, consideremos alguns padrões gerais que podem ser observados em qualquer plotagem residual. Três exemplos aparecem na Figura 12.12. Se a suposição de que a variância de ϵ é idêntica para todos os valores de x, e se o modelo de regressão proposto constituir uma representação adequada da relação entre as variáveis, a plotagem residual deverá dar a impressão geral de uma faixa horizontal de pontos como os do Painel A da Figura 12.12. Entretanto, se a variância de ϵ não for idêntica para todos os valores de x – por exemplo, se a variabilidade nas proximidades da reta de regressão for maior à medida que os valores de x se tornam maiores – um padrão como o do Painel B da Figura 12.12 poderá ser observado. Nesse caso, a hipótese de uma variância constante de ϵ é desrespeitada. Outra plotagem residual possível é mostrada no Painel C. Assim, concluiríamos que o modelo de regressão proposto não é uma representação adequada da relação entre as variáveis. Um modelo de regressão curvilínea ou um modelo de regressão múltipla deve ser considerado.

Retornemos agora à plotagem residual dos restaurantes Armand's Pizza Parlors mostrada na Figura 12.11. Os resíduos parecem aproximar-se do padrão horizontal do Painel A da Figura 12.12. Portanto, concluímos que a plotagem residual não nos fornece evidências de que as suposições feitas sobre o modelo de regressão do Armand's devam ser contestadas.

A essa altura, estamos confiantes na conclusão de que o modelo de regressão linear simples para os restaurantes Armand's é válido.

Figura 12.11 Plotagem dos resíduos em relação à variável independente x para os restaurantes Armand's Pizza Parlors

Experiência e bom julgamento são sempre fatores importantes a serem considerados na interpretação eficiente das plotagens residuais. Raramente uma plotagem residual se molda de maneira precisa a um dos padrões apresentados na Figura 12.12. Contudo, analistas que realizam estudos de regressão com freqüência e revisam plotagens residuais repetidamente tornam-se especialistas em entender as diferenças entre os padrões que são razoáveis e os que indicam que as suposições do modelo devam ser questionadas. Uma plotagem residual constitui uma técnica para avaliar a validade das suposições de um modelo de regressão.

Plotagem Residual em Relação a $\hat{y}$

Outra plotagem residual representa o valor residual da variável dependente $\hat{y}$ no eixo horizontal e os valores residuais no eixo vertical. Para cada resíduo é marcado um ponto no gráfico. A primeira coordenada de cada ponto é dada por $\hat{y}_i$ e a segunda coordenada é dada pelo valor correspondente do i-ésimo resíduo $y_i - \hat{y}_i$. Com os dados do Armand's da Tabela 12.7, as coordenadas do primeiro ponto são (70, –12), correspondentes a $\hat{y}_1 = 70$ e $y_1 - \hat{y}_1 = -12$; as coordenadas do segundo ponto são (90, 15) e assim por diante. A Figura 12.13 apresenta a plotagem residual. Observe que o padrão dessa plotagem residual é idêntico ao padrão da plotagem residual em relação à variável independente x. Esse não é um padrão que nos levaria a questionar as suposições do modelo. Para a regressão linear simples, tanto a plotagem residual e relação a x como a plotagem residual em relação a $\hat{y}$ fornecem o mesmo padrão. Para a análise de regressão múltipla, a plotagem residual em relação a $\hat{y}$ é mais amplamente usada em virtude da presença de mais de uma variável independente.

Figura 12.12 Plotagens residuais de três estudos de regressão

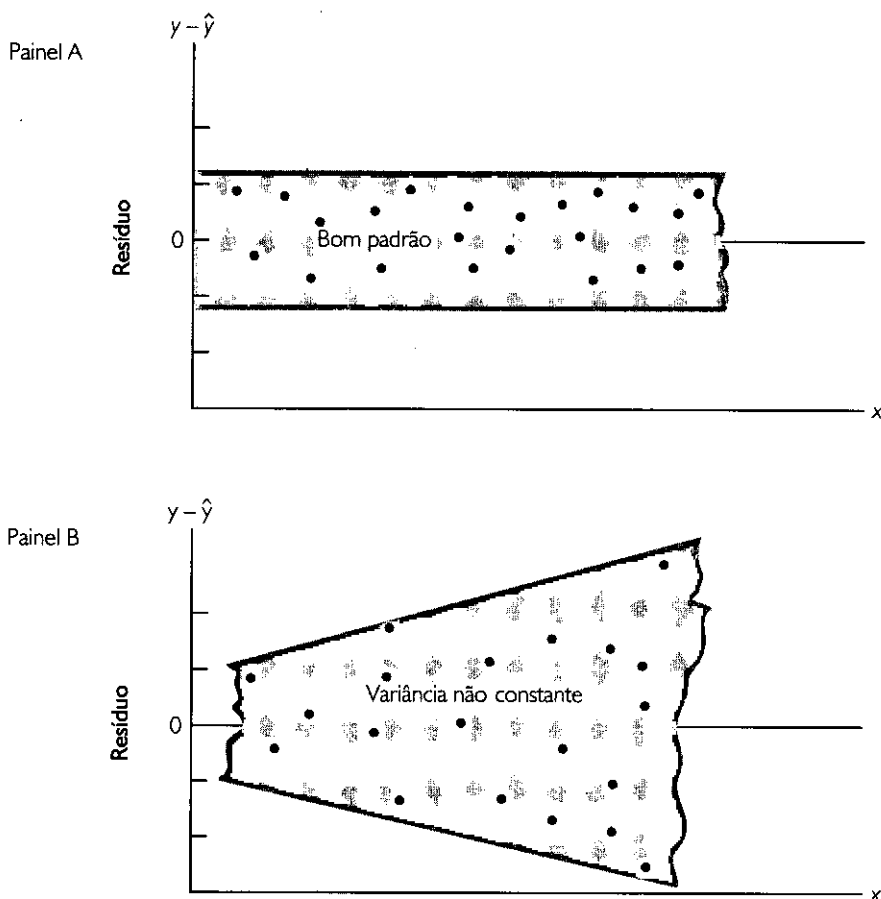



Figura 12.12 Plotagens residuais de três estudos de regressão

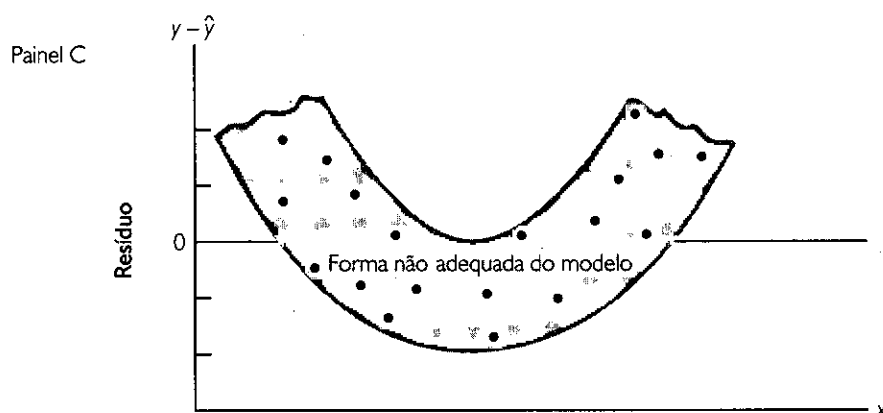
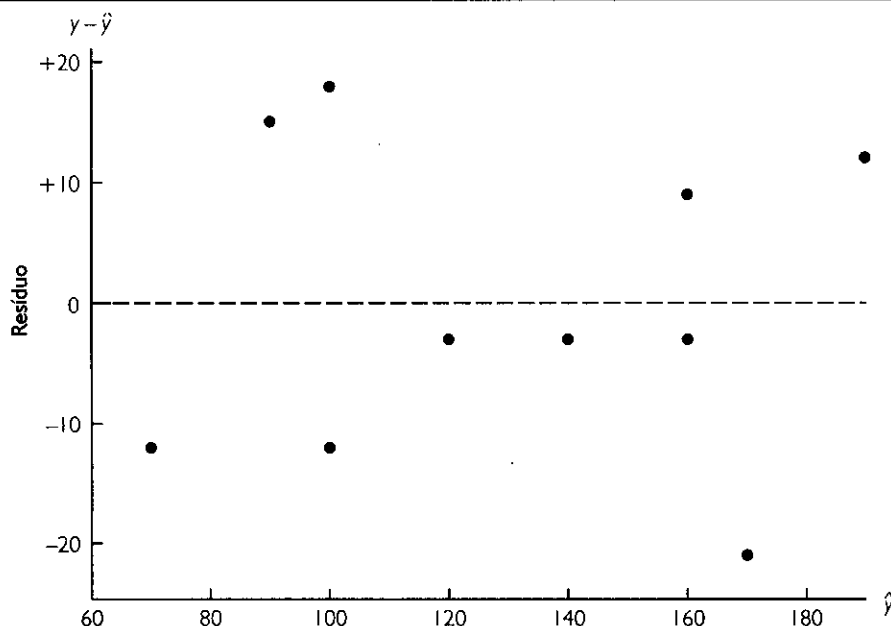



Figura 12.13 Plotagem dos resíduos em relação aos valores de $\hat{y}$ previstos para os restaurantes Armand's Pizza Parlors

NOTAS E COMENTÁRIOS

1. Utilizamos plotagens residuais para validar as suposições de um modelo de regressão. Se nossa revisão indicar que uma ou mais suposições são questionáveis, um modelo de regressão diferente ou uma transformação dos dados devem ser considerados. As medidas corretivas apropriadas quando as suposições são desrespeitadas devem basear-se no bom julgamento; recomendações obtidas de um estatístico experiente podem ser valiosas.
2. A análise de resíduos é o principal método que os estatísticos usam para verificar se as suposições associadas a um modelo de regressão são válidas. Mesmo que nenhuma infração seja encontrada, não decorre necessariamente que o modelo produzirá boas previsões. Entretanto, se testes estatísticos adicionais sustentarem a conclusão de significância e se o coeficiente de determinação for grande, seremos capazes de desenvolver boas estimativas e previsões usando a equação de regressão estimada.

Exercícios

Métodos

AUTOTESTE

45. São dados os valores de duas variáveis, x e y:

x_i	6	11	15	18	20
y_i	6	8	12	20	30

a. Desenvolva uma equação de regressão estimada desses dados.
b. Calcule os resíduos.
c. Desenvolva a plotagem dos resíduos em relação à variável independente x. As suposições sobre os termos de erro parecem ter sido cumpridas?

46. Os dados a seguir foram usados em um estudo de regressão:

Observação	x_i	y_i	Observação	x_i	y_i
1	2	4	6	7	6
2	3	5	7	7	9
3	4	4	8	8	5
4	5	6	9	9	11
5	7	4			

a. Desenvolva uma equação de regressão estimada desses dados.
b. Construa uma plotagem dos resíduos. As suposições a respeito do termo de erro parecem ter sido cumpridas?

Aplicações

AUTOTESTE

47. Dados sobre os dispêndios de publicidade e a receita (em milhares de dólares) do Four Seasons Restaurant são apresentados a seguir:

Dispêndios de Publicidade	Receita
1	19
2	32
4	44
6	40
10	52
14	53
20	54

a. Digamos que x sejam os dispêndios de publicidade e y seja o volume. Use o método dos mínimos quadrados para desenvolver uma aproximação em linha reta da relação entre as duas variáveis.
b. Teste se a receita e os dispêndios de publicidade estão relacionados a um nível de significância de 0,05.
c. Construa uma plotagem residual em relação à variável independente.
d. Quais conclusões você é capaz de tirar da análise residual? Esse modelo deve ser usado ou um modelo melhor deve ser procurado?

48. Consulte o exercício 9, em que uma equação de regressão estimada relacionando os anos de experiência profissional e as vendas anuais foi desenvolvida.

a. Calcule os resíduos e construa uma plotagem residual desse problema.
b. As suposições a respeito do termo de erro parecem razoáveis em função da plotagem residual?

49. Os American Depository Receipts (ADRs) são certificados negociados na Bolsa de Valores de Nova York (Nyse) representando ações de uma empresa estrangeira, as quais são mantidas em depósito em um banco de seu país de origem. A tabela a seguir exibe a razão preço/rendimentos (P/R) e o retorno percentual do investimento (ROI) correspondentes a dez empresas indianas que provavelmente são novas ADRs (*Bloomberg Personal Finance*, abril de 2000).

ARQUIVO DA INTERNET
ADRs

Empresa	Retorno do Investimento (ROI)	Razão Preço/ Rendimento (P/R)
Bharti Televentures	6,43	36,88
Gujarat Ambuja Cements	13,49	27,03
Hindalco Industries	14,04	10,83
ICICI	20,67	5,15
Mahanagar Telephone Nigam	22,74	13,35
NIIT	46,23	95,59

Empresa	Retorno do Investimento (ROI)	Razão Preço/ Rendimento (P/R)
Pentamedia Graphics	28,90	54,85
Satyam Computer Services	54,01	189,21
Silverline Technologies	28,02	75,86
Videsh Sanchar Nigam	27,04	13,17

a. Use um software para desenvolver uma equação de regressão estimada relacionando y = P/R e x = ROI.
b. Construa uma plotagem residual em relação à variável independente.
c. As suposições a respeito dos termos de erro e da forma do modelo parecem razoáveis em função da plotagem residual?

Resumo

Neste capítulo, mostramos como a análise de regressão pode ser usada para determinar como uma variável dependente y se relaciona com uma variável independente x. Na regressão linear simples, o modelo de regressão é $y = \beta_0 + \beta_1 x + \epsilon$. A equação de regressão linear simples $E(y) = \beta_0 + \beta_1 x$ descreve como o valor médio ou esperado de y está relacionado a x. Utilizamos dados amostrais e o método dos mínimos quadrados para desenvolver a equação de regressão estimada $\hat{y} = b_0 + b_1 x$. Com efeito, b_0 e b_1 são as estatísticas amostrais usadas para estimar os parâmetros desconhecidos do modelo, β_0 e β_1.

O coeficiente de determinação foi apresentado como uma medida da eficiência de ajuste da equação de regressão estimada; ele pode ser interpretado como a proporção da variação na variável dependente y que pode ser explicada pela equação de regressão estimada. Revisamos a correlação como uma medida descritiva de uma relação linear entre duas variáveis.

As suposições acerca do modelo de regressão e seu termo de erro ϵ associado foram discutidos, e os testes t e F, baseados nessas suposições, foram apresentados como um meio de determinar se a relação entre duas variáveis é estatisticamente significativa. Mostramos como usar a equação de regressão estimada para desenvolver estimações por intervalo de confiança do valor médio de y e as estimações por intervalo de previsão de valores individuais de y.

Este capítulo se encerrou com uma seção sobre a solução computadorizada dos problemas de regressão, e uma seção sobre o uso da análise residual para validar as suposições do modelo.

Glossário

Variável dependente A variável que está sendo prevista. É designada y.

Variável independente A variável que é usada para prever o valor da variável independente. É designada x.

Regressão linear simples Uma análise de regressão que envolve uma variável independente e uma variável dependente, na qual a relação entre as variáveis é aproximada por uma linha reta.

Modelo de regressão A equação que descreve como y está relacionado com x e um termo de erro; na regressão linear simples, o modelo de regressão é $y = \beta_0 + \beta_1 x + \epsilon$.

Equação de regressão A equação que descreve como a média, ou valor esperado da variável dependente, está relacionada com a variável independente; na regressão linear simples, $E(y) = \beta_0 + \beta_1 x$.

Equação de regressão estimada A estimativa da equação de regressão desenvolvida a partir de dados amostrais usando-se o método dos mínimos quadrados. Para a regressão linear simples, a equação de regressão estimada é $\hat{y} = b_0 + b_1 x$.

Método dos mínimos quadrados Um procedimento para se usar dados amostrais com a finalidade de encontrar a equação de regressão estimada. O objetivo é minimizar $\Sigma(y_i - \hat{y}_i)^2$.

Diagrama de dispersão Um gráfico de dados bivariáveis no qual a variável independente se situa no eixo horizontal e a variável dependente, no eixo vertical.

Coeficiente de determinação Uma medida da eficiência do ajuste da equação de regressão estimada. Pode ser interpretado como a proporção da variabilidade da variável dependente y que é explicada pela equação de regressão estimada.

***i*-ésimo resíduo** A diferença entre o valor observado da variável dependente e o valor previsto usando-se a equação de regressão estimada; para a i-ésima observação, o i-ésimo resíduo é $y_i - \hat{y}_i$.

Coeficiente de correlação Uma medida da intensidade da relação linear entre duas variáveis (discutido anteriormente, no Capítulo 3).

Erro médio quadrático A estimativa sem viés da variância do termo de erro σ^2. É designado MSE ou s^2.

Erro padrão da estimativa A raiz quadrada da média do erro médio quadrático, designado s. É a estimativa de σ, que é o desvio padrão do termo de erro ϵ.

Tabela ANOVA A tabela da análise de variância usada para resumir os cálculos associados ao teste F de significância.

Intervalo de confiança A estimação por intervalo do valor médio de y para determinado valor de x.

Intervalo de previsão A estimação por intervalo de um valor individual de y para determinado valor de x.

Análise residual A principal ferramenta para determinar se o modelo de regressão proposto é apropriado.

Plotagem residual Representação gráfica dos resíduos que pode ser usada para determinar se as suposições feitas a respeito do modelo de regressão parecem ser válidas.

Fórmulas-Chave

Modelo de Regressão Linear Simples

$$y = \beta_0 + \beta_1 x + \epsilon \tag{12.1}$$

Equação de Regressão Linear Simples

$$E(y) = \beta_0 + \beta_1 x \tag{12.2}$$

Equação de Regressão Linear Simples Estimada

$$\hat{y} = b_0 + b_1 x \tag{12.3}$$

Critério dos Mínimos Quadrados

$$\min \Sigma(y_i - \hat{y}_i)^2 \tag{12.5}$$

Inclinação e Interseção com *y* na Equação de Regressão Estimada

$$b_1 = \frac{\Sigma(x_i - \bar{x})(y_i - \bar{y})}{\Sigma(x_i - \bar{x})^2} \tag{12.6}$$

$$b_0 = \bar{y} - b_1 \bar{x} \tag{12.7}$$

Soma dos Quadrados dos Erros

$$\text{SSE} = \Sigma(y_i - \hat{y}_i)^2 \tag{12.8}$$

Soma Total dos Quadrados

$$\text{SST} = \Sigma(y_i - \bar{y})^2 \tag{12.9}$$

Soma dos Quadrados da Regressão

$$\text{SSR} = \Sigma(\hat{y}_i - \bar{y})^2 \tag{12.10}$$

Relação entre SST, SSR e SSE

$$\text{SST} = \text{SSR} + \text{SSE} \tag{12.11}$$

Coeficiente de Determinação

$$r^2 = \frac{\text{SSR}}{\text{SST}} \tag{12.12}$$

Coeficiente de Correlação da Amostra

$$\begin{aligned} r_{xy} &= (\text{sinal de } b_1)\sqrt{\text{Coeficiente de determinação}} \\ &= (\text{sinal de b}_1)\sqrt{\text{r}^2} \end{aligned} \tag{12.13}$$

Erro Médio Quadrático (Estimativa de σ^2)

$$s^2 = \text{MSE} = \frac{\text{SSE}}{n - 2} \tag{12.15}$$

Erro Padrão da Estimativa

$$s = \sqrt{\text{MSE}} = \sqrt{\frac{\text{SSE}}{n-2}} \quad (12.16)$$

Desvio Padrão de b_1

$$\sigma_{b_1} = \frac{\sigma}{\sqrt{\Sigma(x_i - \bar{x})^2}} \quad (12.17)$$

Desvio Padrão Estimado de b_1

$$s_{b_1} = \frac{s}{\sqrt{\Sigma(x_i - \bar{x})^2}} \quad (12.18)$$

Estatística de Teste t

$$t = \frac{b_1}{s_{b_1}} \quad (12.19)$$

Regressão pela Média Quadrática

$$\text{MSR} = \frac{\text{SSR}}{\text{Números de variáveis independentes}} \quad (12.20)$$

Estatística de Teste F

$$F = \frac{\text{MSR}}{\text{MSE}} \quad (12.21)$$

Desvio Padrão Estimado de $\hat{y}_p$

$$s_{\hat{y}_p} = s\sqrt{\frac{1}{n} + \frac{(x_p - \bar{x})^2}{\Sigma(x_i - \bar{x})^2}} \quad (12.23)$$

Intervalo de Confiança de $E(y_p)$

$$\hat{y}_p \pm t_{\alpha/2} s_{\hat{y}_p} \quad (12.24)$$

Desvio Padrão Estimado de um Valor Individual

$$s_{\text{ind}} = s\sqrt{1 + \frac{1}{n} + \frac{(x_p - \bar{x})^2}{\Sigma(x_i - \bar{x})^2}} \quad (12.26)$$

Intervalo de Previsão de y_p

$$\hat{y}_p \pm t_{\alpha/2} s_{\text{ind}} \quad (12.27)$$

Resíduo da Observação i

$$y_i - \hat{y}_i \quad (12.28)$$

Exercícios Suplementares

50. Os dados apresentados na tabela a seguir exibem o número de vendas de ações (em milhões) e o preço esperado (média do menor preço projetado e do maior preço projetado) de dez ofertas selecionadas de oferta pública inicial de títulos.

Empresa	Venda de Ações	Preço Esperado (US$)
American Physician	5,0	15
Apex Silver Mines	9,0	14
Dan River	6,7	15
Franchise Mortgage	8,75	17
Gene Logic	3,0	11
International Home Foods	13,6	19
PRT Group	4,6	13
Rayovac	6,7	14
RealNetworks	3,0	10
Software AG Systems	7,7	13

ARQUIVO DA INTERNET
IPO

a. Desenvolva uma equação de regressão estimada, sendo o número de vendas de ações a variável independente e o preço esperado a variável dependente.
b. No nível de significância de 0,05, há uma relação significativa entre as duas variáveis?
c. A equação de regressão estimada proporcionou um bom ajuste? Explique.
d. Use a equação de regressão estimada para estimar o preço esperado por uma empresa que considera uma oferta pública inicial de 6 milhões de ações?

51. Os programas corporativos de recompra de ações freqüentemente são propalados como um benefício para os acionistas. Robert Gabele, diretor de pesquisa interna do First Call/Thomson Financial, notou que muitos desses programas são levados a efeito com o único intuito de adquirirem ações das *incentive options* de uma empresa para sua alta gerência. Em todas as empresas, as *stock options*[8] existentes em 1998 representavam 6,2% de todas as ações ordinárias em circulação. Os dados a seguir mostram o número de ações cobertas pelas *option grants* e o número de ações ordinárias em circulação de 13 empresas (*Bloomberg Personal Finance*, janeiro/fevereiro de 2000).

ARQUIVO DA INTERNET
Options

Empresa	Número de Ações das *Option Grants* em Circulação (milhões)	Ações Ordinárias em Circulação (milhões)
Adobe Systems	20,3	61,8
Apple Computer	52,7	160,9
Applied Materials	109,1	375,4
Autodesk	15,7	58,9
Best Buy	44,2	203,8
Fruit of the Loom	14,2	66,9
ITT Industries	18,0	87,9
Merrill Lynch	89,9	365,5
Novell	120,2	335,0
Parametric Technology	78,3	269,3
Reebok International	12,8	56,1
Silicon Graphics	52,6	188,8
Toys R Us	54,8	247,6

a. Desenvolva a equação de regressão estimada que possa ser usada para estimar o número de ações das *option grants* em circulação, dado o número de ações ordinárias em circulação.
b. Use a equação de regressão estimada para estimar o número de ações das *option grants* em circulação de uma empresa que tem 150 milhões de ações ordinárias em circulação.
c. Você acredita que a equação de regressão estimada forneça uma boa previsão do número de ações das *option grants* em circulação? Use r^2 para sustentar sua resposta.

52. A *Bloomberg Personal Finance* (julho/agosto de 2001) publicou que o título com beta[9] da Texas Instruments era de 1,46. Os títulos com beta para títulos individuais são determinados por regressão linear simples. Para cada título, a variável dependente é o seu retorno percentual trimestral (valorização do capital mais os dividendos) menos seu retorno percentual trimestral que possa ser obtido de um investimento isento de riscos (a taxa de Letras do Tesouro Nacional é usada como o índice isento de riscos). A variável independente é o retorno percentual trimestral (valorização do capital mais os dividendos) do mercado financeiro (S&P 500) menos o retorno percentual de um investimento isento de riscos. Uma equação de regressão estimada é desenvolvida com os dados trimestrais; o título com beta é a inclinação (declive) da equação de regressão estimada (b_1). O valor do título com beta muitas vezes é interpretado como uma medida do risco associado ao título. Títulos com beta maiores que 1 indicam que o título é mais volátil que a média do mercado; títulos com beta menores que 1 indicam que o título é menos volátil que a média do mercado. Suponha que os seguintes valores sejam as diferenças entre o retorno percentual e o rendimento isento de riscos de dez trimestres para a S&P 500 e a Horizon Technology.

---

[8] NT: *Stock Option* – A empresa oferece aos funcionários opções de compra de suas ações. Esse benefício está atrelado ao desempenho, ao cumprimento de metas. Por exemplo, membros da diretoria podem subscrever ações em um momento determinado e com preço inferior ao estimado pelo mercado. É uma forma de motivar o pessoal. *Option grant* é o mesmo que *option stock grants* e se refere a esse tipo de oferta de ações.

[9] NT: Título com beta – Medida de risco diversificável de um ativo. Coeficiente de risco de mercado da carteira durante o período analisado.

S&P 500	Horizon
1,2	−0,7
−2,5	−2,0
−3,0	−5,5
2,0	4,7
5,0	1,8
1,2	4,1
3,0	2,6
−1,0	2,0
,5	−1,3
2,5	5,5

a. Desenvolva uma equação de regressão estimada que possa ser usada para determinar o título com beta da Horizon Technology. Qual é o título com beta da Horizon Technology?
b. Teste se há uma relação significativa no nível de significância de 0,05.
c. A equação de regressão estimada proporcionou um bom ajuste? Explique.
d. Use os títulos com beta da Texas Instruments e da Horizon Technology para comparar o risco associado aos dois títulos financeiros.

53. O State of the Service Report 2002-2003 da Australian Public Service Commission divulgou as avaliações de satisfação no trabalho dos empregados. Uma das questões da pesquisa pediu aos empregados que escolhessem os cinco fatores (de uma lista de fatores) mais importantes no ambiente de trabalho que afetavam mais fortemente o quanto estavam satisfeitos no emprego. Os entrevistados foram solicitados a indicar o nível de satisfação correspondente aos cinco fatores importantes por eles indicados. Os dados a seguir apresentam a porcentagem de empregados que citaram determinado fator entre os cinco principais e o correspondente nível de satisfação, medido em termos da porcentagem de empregados que citaram o fator entre os cinco principais, e que estavam "muito satisfeitos" ou "satisfeitos" com esse fator em seus ambientes de trabalho atuais (http://www.apsc.gov.au/stateoftheservices).

Fator do Ambiente de Trabalho	Os Cinco Principais (%)	Avaliação de Satisfação (%)
Carga de trabalho apropriada	30	49
Oportunidade para ser criativo(a)/inovador(a)	38	64
Oportunidade de dar uma contribuição útil à sociedade	40	67
Deveres/Expectativas definidas claramente	40	69
Programas de trabalho flexíveis	55	86
Boas relações trabalhistas	60	85
Oferta de tarefas interessantes	48	74
Oportunidades de desenvolvimento da carreira	33	43
Oportunidades para desenvolver novas habilidades	46	66
Oportunidades para utilizar minhas habilidades	50	70
Retorno e reconhecimento habituais do esforço	42	53
Salário	47	62
Ver resultados palpáveis do meu trabalho	42	69

ARQUIVO DA INTERNET
JobSat

a. Desenvolva um diagrama de dispersão com os Cinco Principais (%) no eixo horizontal e a Avaliação de Satisfação (%) no eixo vertical.
b. O que o diagrama de dispersão desenvolvido no item (a) indica a respeito da relação entre as duas variáveis?
c. Desenvolva uma equação de regressão estimada que possa ser usada para prever a Avaliação de Satisfação (%), dados os Cinco Principais (%).
d. Teste se há uma relação significativa ao nível de significância de 0,05.
e. A equação de regressão estimada proporcionou um bom ajuste? Explique.
f. Qual é o valor do coeficiente de correlação da amostra?

54. A Jensen Tire & Auto está em vias de decidir se assina ou não um contrato de manutenção de seu novo equipamento computadorizado de alinhamento e balanceamento de pneus. Os gerentes acham que as despesas de manutenção devem relacionar-se com o uso, e coletaram as seguintes informações sobre o uso semanal (em horas) e as despesas anuais de manutenção (em centenas de dólares).

Horas de Uso Semanal	Despesas Anual de Manutenção
13	17,0
10	22,0
20	30,0
28	37,0
32	47,0
17	30,5
24	32,5
31	39,0
40	51,5
38	40,0

a. Desenvolva a equação de regressão estimada que relacione a despesa de manutenção com o uso semanal.
b. Teste a significância da relação do item (a) ao nível de significância 0,05.
c. A Jensen espera usar o novo equipamento 30 horas por semana. Desenvolva um intervalo de previsão de 95% relativo à despesa anual de manutenção da empresa.
d. Se o contrato de manutenção custar US$ 3 mil por ano, você recomendaria adquiri-lo? Por quê?

55. Em um processo de manufatura, avaliou-se que a velocidade da linha de montagem (metros por minuto) afeta o número de peças defeituosas encontradas durante o processo de inspeção. Para testar essa teoria, os gerentes idealizaram uma situação na qual o mesmo lote de peças era inspecionado visualmente em diversas velocidades da linha de montagem. Coletaram os seguintes dados.

Velocidade da linha de produção (em metros/min)	Número de peças defeituosas encontradas
6	21
6	19
12	15
9	16
18	14
12	17

a. Desenvolva a equação de regressão estimada que relacione a velocidade da linha de montagem com o número de peças defeituosas encontradas.
b. No nível de significância 0,05, determine se a velocidade da linha de montagem e o número de peças defeituosas encontradas se relacionam.
c. A equação de regressão estimada proporciona um bom ajuste aos dados?
d. Desenvolva um intervalo de confiança para prever o número médio de peças defeituosas para uma linha de montagem cuja velocidade é de 15,24 m por minuto.

56. Um sociólogo foi contratado pelo hospital de uma grande cidade para investigar a relação entre o número de dias não-autorizados em que os funcionários se ausentavam do trabalho por ano e a distância (em quilômetros) entre a casa e o trabalho dos empregados. Uma amostra de dez empregados foi escolhida, e os seguintes dados foram coletados:

Distância do trabalho (km)	Número de Dias Ausentes
1,6	8
4,8	5
6,4	8
9,6	7
12,86	6
16,1	3
19,3	5
22,5	2
22,5	4
29,0	2

a. Desenvolva um diagrama de dispersão desses dados. Uma relação linear parece razoável? Explique.
b. Desenvolva a equação de regressão estimada pelo método dos mínimos quadrados.
c. Há uma relação significativa entre as duas variáveis? Use $\alpha = 0{,}05$.

d. A equação de regressão estimada proporcionou um bom ajuste? Explique.

e. Use a equação de regressão estimada desenvolvida no item (b) para desenvolver um intervalo de confiança de 95% do número esperado de dias que os empregados que moram a 8 km da empresa se ausentarão do trabalho.

57. O departamento regional de trânsito de uma grande região metropolitana quer determinar se há alguma relação entre a idade de um ônibus e o custo anual de manutenção. Uma amostra de dez ônibus resultou nos seguintes dados:

Idade do Ônibus (anos)	Custo de Manutenção (US$)
1	350
2	370
2	480
2	520
2	590
3	550
4	750
4	800
5	790
5	950

a. Desenvolva a equação de regressão estimada pelo método dos mínimos quadrados.

b. Teste se as duas variáveis são significativamente relacionadas com $\alpha = 0{,}05$.

c. A reta dos mínimos quadrados proporcionou um bom ajuste para os dados observados? Explique.

d. Desenvolva um intervalo de previsão de 96% para o custo de manutenção de um ônibus específico que tem 4 anos.

58. Um professor de Marketing do Givens College está interessado na relação entre as horas que um aluno gasta estudando e a pontuação total obtida em um curso. Dados coletados sobre dez estudantes que fizeram o curso no último trimestre são os seguintes:

Horas que o Aluno Gasta Estudando	Pontuação Total Obtida
45	40
30	35
90	75
60	65
105	90
65	50
90	90
80	80
55	45
75	65

a. Desenvolva uma equação de regressão estimada mostrando como a pontuação total obtida se relaciona com as horas que o aluno gasta estudando.

b. Teste a significância do modelo com $\alpha = 0{,}05$.

c. Preveja a pontuação total obtida por Mark Sweeney. Ele passou 95 horas estudando.

d. Desenvolva um intervalo de previsão de 95% da pontuação total obtida por Mark Sweeney.

59. A Transactional Records Access Clearinghouse, da Syracuse University, divulgou dados mostrando as disparidades em uma auditoria realizada pelo Internal Revenue Service (Departamento da Receita Federal). A tabela a seguir apresenta a média da renda bruta ajustada declarada e a porcentagem das declarações auditadas em 20 áreas do IRS selecionadas.

Área	Renda Bruta Ajustada (US$)	Porcentagem de Declarações Auditadas
Los Angeles	36.664	1,3
Sacramento	38.845	1,1
Atlanta	34.886	1,1
Boise	32.512	1,1
Dallas	34.531	1,0
Providence	35.995	1,0
São José	37.799	0,9
Cheyenne	33.876	0,9
Fargo	30.513	0,9

ARQUIVO DA INTERNET

IRSAudit

Renda Bruta Área	Porcentagem de Ajustada (US$)	Declarações Auditadas
New Orleans	30.174	0,9
Oklahoma City	30.060	0,8
Houston	37.153	0,8
Portland	34.918	0,7
Phoenix	33.291	0,7
Augusta	31.504	0,7
Albuquerque	29.199	0,6
Greensboro	33.072	0,6
Columbia	30.859	0,5
Nashville	32.566	0,5
Buffalo	34.296	0,5

a. Desenvolva a equação de regressão estimada que possa ser usada para prever a porcentagem de declarações auditadas, dada a média de renda bruta ajustada declarada.

b. Ao nível de 0,05 de significância determine se a renda bruta ajustada e a porcentagem de declarações auditadas se relacionam.

c. Use a equação de regressão estimada desenvolvida no item (a) para calcular um intervalo de confiança de 95% para a porcentagem de declarações auditadas correspondentes às áreas com uma média de renda bruta ajustada de US$ 35 mil.

Estudo de Caso I – Gastos e Desempenho Escolar

O nível de progresso educacional dos estudantes está relacionado com os investimentos que o Estado em que eles residem faz em educação? Em muitas comunidades, os contribuintes fazem essa importante pergunta, uma vez que os distritos escolares solicitam aumentos da parcela do imposto de renda destinada à educação. Nesse caso, você será solicitado a analisar dados sobre os gastos e sobre as notas de desempenho estudantil a fim de determinar se há alguma relação entre os gastos e o desempenho estudantil nas escolas públicas.

O programa *National Assessment of Educational Progress* – Avaliação Nacional do Progresso Escolar (NAEP) do governo federal norte-americano é usado freqüentemente para medir o progresso escolar dos estudantes. A Tabela 12.8 mostra o atual custo total anual por aluno e a pontuação NAEP média correspondente aos 35 estados que participaram do programa. No site www.thomsonlearning.com.br/estatapl.htm há um arquivo intitulado NAEP. A pontuação média do exame é a soma das notas obtidas em matemática, ciências e leitura no teste NAEP de 1996 (1994 para leitura).

ARQUIVO DA INTERNET NAEP

Tabela 12.8 Gastos por aluno e pontuação média nos estados que participaram do programa NAEP

Estado	Gastos por Aluno (US$)	Pontuação Média
Louisiana	4.049	581
Mississippi	3.423	582
Califórnia	4.917	580
Havaí	5.532	580
Carolina do Sul	4.304	603
Alabama	3.777	604
Georgia	4.663	611
Flórida	4.934	611
Novo México	4.097	614
Arkansas	4.060	615
Delaware	6.208	615
Tennessee	3.800	618
Arizona	4.041	618
West Virginia	5.247	625
Maryland	6.100	625
Kentucky	5.020	626
Texas	4.520	627

Tabela 12.8 Gastos por aluno e pontuação média nos estados que participaram do programa NAEP (continuação)

Estado	Gastos por Aluno (US$)	Pontuação Média
Nova York	8.162	628
Carolina do Norte	4.521	629
Rhode Island	6.554	638
Washington	5.338	639
Missouri	4.483	641
Colorado	4.772	644
Indiana	5.128	649
Utah	3.280	650
Wyoming	5.515	657
Connecticut	7.629	657
Massachusetts	6.413	658
Nebraska	5.410	660
Minnesota	5.477	661
Iowa	5.060	665
Montana	4.985	667
Wisconsin	6.055	667
Dakota do Norte	4.374	671
Maine	5.561	675

Os alunos avaliados estão na oitava série, exceto os do exame de leitura, que é aplicado apenas a alunos da quarta série. A pontuação máxima possível é 1.300. A Tabela 12.9 apresenta os gastos por aluno em 13 estados que não participaram de pesquisas relevantes do programa NAEP. Esses dados foram publicados em um artigo sobre o nível de gastos e o desempenho escolar na revista *Forbes* (3 de novembro de 1997).

Relatório Administrativo

1. Desenvolva resumos numéricos e gráficos dos dados.
2. Use análise de regressão para investigar a relação entre a quantia gasta por aluno e a pontuação média no exame NAEP. Discuta suas conclusões.

Tabela 12.9 Gastos por aluno nos estados que não participaram do programa NAEP

Estado	Gasto por Aluno (US$)
Idaho	3.602
Dakota do Sul	4.067
Oklahoma	4.265
Nevada	4.658
Kansas	5.164
Illinois	5.297
New Hampshire	5.387
Ohio	5.438
Oregon	5.588
Vermont	6.269
Michigan	6.391
Pennsylvania	6.579
Alaska	7.890

3. Você acha que a equação de regressão estimada desenvolvida para esses dados poderia ser usada para estimar as pontuações médias de exame nos estados que não participaram do programa NAEP?
4. Suponha que você tenha considerado somente os estados que gastam no mínimo US$ 4 mil, mas não mais do que US$ 6 mil por aluno. Quanto a esses estados, a relação entre as duas variáveis parece ser de alguma maneira diferente do conjunto de dados completo? Discuta os resultados de suas conclusões e se você acha apropriado excluir os estados que gastam menos de US$ 4 mil e mais de US$ 6 mil por aluno ao ano.

5. Desenvolva estimativas da média de pontuações obtidas nos estados que não participaram no programa NAEP.
6. Com base em suas análises, você acha que o nível de progresso escolar dos estudantes se relaciona com a quantidade de investimentos que o estado faz em educação?

Estudo de Caso 2 – U.S. Department of Transportation

Como parte de um estudo sobre segurança no trânsito, o U.S. Department of Transportation (Departamento de Transportes dos Estados Unidos) coletou dados sobre o número de acidentes fatais para cada mil carteiras de habilitação, bem como a porcentagem de motoristas com menos de 21 anos autorizados a dirigir, em uma amostra de 42 cidades. Os dados coletados ao longo de um ano são apresentados a seguir. Esses dados estão disponíveis no site www.thomsonlearning.com.br/estatapl.htm, no arquivo intitulado Safety.

ARQUIVO DA INTERNET
Safety

Porcentagem com Menos de 21	Acidentes Fatais para Cada 1000 Carteiras de Habilitação	Porcentagem com Menos de 21	Acidentes Fatais para Cada 1000 Carteiras de Habilitação
13	2,962	17	4,100
12	0,708	8	2,190
8	0,885	16	3,623
12	1,652	15	2,623
11	2,091	9	0,835
17	2,627	8	0,820
18	3,830	14	2,890
8	0,368	8	1,267
13	1,142	15	3,224
8	0,645	10	1,014
9	1,028	10	0,493
16	2,801	14	1,443
12	1,405	18	3,614
9	1,433	10	1,926
10	0,039	14	1,643
9	0,338	16	2,943
11	1,849	12	1,913
12	2,246	15	2,814
14	2,855	13	2,634
14	2,352	9	0,926
11	1,294	17	3,256

Relatório Administrativo

1. Desenvolva resumos numéricos e gráficos dos dados.
2. Use análise de regressão para investigar a relação entre o número de acidentes fatais e a porcentagem de motoristas com menos de 21 anos. Discuta suas conclusões.
3. Qual conclusão e quais recomendações você é capaz de deduzir de sua análise?

Estudo de Caso 3 – Doações de Ex-Alunos

As doações de ex-alunos são uma fonte importante de receita para colégios e universidades. Se os administradores pudessem determinar os fatores que influem no aumento da porcentagem de ex-alunos que fazem doações, talvez pudessem ser capazes de implementar políticas que levassem a um aumento das receitas. Pesquisas mostram que os estudantes que estão mais satisfeitos em seus contatos com os professores têm mais probabilidade de graduar-se. Em conseqüência, poder-se-ia imaginar que classes menores, e uma razão professor/alunos, acarretariam uma maior porcentagem de graduados satisfeitos, o que, por sua vez, poderia levar a um aumento na porcentagem de ex-alunos que fazem doações. A Tabela 12.10 apresenta dados de 48 universidades federais (*America's Best Colleges*, edição de 2000). A coluna intitu-

lada Porcentagem de Classes com Menos de 20 exibe a porcentagem de classes disponíveis com menos de 20 alunos. A coluna intitulada Razão Estudantes/Professor é o número de estudantes matriculados dividido pelo número total de professores. Finalmente, a coluna intitulada Índice de Doação de Ex-alunos é a porcentagem de ex-alunos que fizeram doações à universidade.

Relatório Administrativo

1. Desenvolva resumos numéricos e gráficos dos dados.
2. Use análise de regressão para desenvolver uma equação de regressão estimada que possa ser usada para prever o índice de doação de ex-alunos, dada a porcentagem de classes com menos de 20 alunos.
3. Use análise de regressão para desenvolver uma equação de regressão estimada que possa ser usada para prever o índice de doação de ex-alunos, dada a relação estudantes/professores.
4. Qual das duas equações de regressão estimadas proporciona o melhor ajuste? Quanto a essa equação de regressão estimada, realize uma análise dos resíduos e discuta suas descobertas e conclusões.
5. Quais conclusões e recomendações você é capaz de deduzir de sua análise?

Tabela 12.10 Dados de 48 Universidades Federais

	% de Classes com Menos de 20	Razão Estudante/Professor	Índice de Doação de Ex-Alunos
Boston College	39	13	25
Brandeis University	68	8	33
Brown University	60	8	40
California Institute of Technology	65	3	46
Carnegie Mellon University	67	10	28
Case Western Reserve Univ.	52	8	31
College of William and Mary	45	12	27
Columbia University	69	7	31
Cornell University	72	13	35
Dartmouth College	61	10	53
Duke University	68	8	45
Emory University	65	7	37
Georgetown University	54	10	29
Harvard University	73	8	46
Johns Hopkins University	64	9	27
Lehigh University	55	11	40
Massachusetts Inst. of Technology	65	6	44
New York University	63	13	13
Northwestern University	66	8	30
Pennsylvania State Univ.	32	19	21
Princeton University	68	5	67
Rice University	62	8	40
Stanford University	69	7	34
Tufts University	67	9	29
Tulane University	56	12	17
U. of California–Berkeley	58	17	18
U. of California–Davis	32	19	7
U. of California–Irvine	42	20	9
U. of California–Los Angeles	41	18	13
U. of California–San Diego	48	19	8
U. of California–Santa Barbara	45	20	12
U. of Chicago	65	4	36
U. of Florida	31	23	19
U. of Illinois–Urbana Champaign	29	15	23
U. of Michigan–Ann Arbor	51	15	13

ARQUIVO DA INTERNET
Alumni

Tabela 12.10 Dados de 48 Universidades Federais (continuação)

	% de Classes com Menos de 20	Razão Estudante/Professor	Índice de Doação de Ex-alunos
U. of North Carolina–Chapel Hill	40	16	26
U. of Notre Dame	53	13	49
U. of Pennsylvania	65	7	41
U. of Rochester	63	10	23
U. of Southern California	53	13	22
U. of Texas–Austin	39	21	13
U. of Virginia	44	13	28
U. of Washington	37	12	12
U. of Wisconsin–Madison	37	13	13
Vanderbilt University	68	9	31
Wake Forest University	59	11	38
Washington University–St. Louis	73	7	33
Yale University	77	7	50

Estudo de Caso 4 – Valores dos Times de Beisebol da Major League[10]

Um grupo dirigido por John Henry pagou US$ 700 milhões para comprar o Boston Red Sox, não obstante essa equipe não ter ganho a World Series[11] desde 1918, e anunciaram um prejuízo operacional de US$ 11,4 milhões para 2001. Além disso, a revista *Forbes* estima que o valor atual do time é, de fato, US$ 426 milhões. A *Forbes* atribui a diferença entre o valor atual de uma equipe e o preço que os investidores estão dispostos a pagar ao fato de a compra de um time freqüentemente incluir a aquisição de uma rede de TV a cabo flagrantemente subavaliada. Por exemplo, ao comprar o Boston Red Sox, os novos proprietários também adquiriram uma participação de 80% na New England Sports Network. A Tabela 12.11 apresenta os dados das 30 principais equipes da liga (*Forbes,* 15 de abril de 2002). A coluna intitulada Valor contém os valores das equipes baseados nos atuais negócios com estádios, sem dedução de débitos. A coluna intitulada Renda indica os ganhos antes dos juros, dos impostos e da desvalorização.

Relatório Administrativo

1. Desenvolva resumos numéricos e gráficos dos dados.
2. Use análise de regressão para investigar a relação entre valor e renda. Discuta suas conclusões.

ARQUIVO DA INTERNET

Tabela 12.11 Dados referentes aos times de Beisebol da Major League

Time	Valor	Receita	Renda
New York Yankees	730	215	18,7
New York Mets	482	169	14,3
Los Angeles Dodgers	435	143	−29,6
Boston Red Sox	426	152	−11,4
Atlanta Braves	424	160	9,5
Seattle Mariners	373	166	14,1
Cleveland Indians	360	150	−3,6
Texas Rangers	356	134	−6,5
San Francisco Giants	355	142	16,8
Colorado Rockies	347	129	6,7
Houston Astros	337	125	4,1
Baltimore Orioles	319	133	3,2
Chicago Cubs	287	131	7,9

[10] NT: *Major League(s)* – As duas principais ligas de clubes de beisebol nos Estados Unidos: a National League e a American League.
[11] NT: *World Series* – Uma série de jogos anuais entre os times vencedores das duas principais ligas de beisebol dos Estados Unidos para decidir o campeonato.

Tabela 12.11 Dados referentes aos times de Beisebol da Major League (continuação)

Time	Valor	Receita	Renda
Arizona Diamondbacks	280	127	−3,9
St. Louis Cardinals	271	123	−5,1
Detroit Tigers	262	114	12,3
Pittsburgh Pirates	242	108	9,5
Milwaukee Brewers	238	108	18,8
Philadelphia Phillies	231	94	2,6
Chicago White Sox	223	101	−3,8
San Diego Padres	207	92	5,7
Cincinnati Reds	204	87	4,3
Anaheim Angels	195	103	5,7
Toronto Blue Jays	182	91	−20,6
Oakland Athletics	157	90	6,8
Kansas City Royals	152	85	2,2
Tampa Bay Devil Rays	142	92	−6,1
Florida Marlins	137	81	1,4
Minnesota Twins	127	75	3,6
Montreal Expos	108	63	−3,4

3. Use análise de regressão para investigar a relação entre valor e receita. Discuta suas conclusões.
4. Quais conclusões e recomendações você é capaz de deduzir de sua análise?

Apêndice 12.1 – Análise de Regressão com o Minitab

Na Seção 12.7, discutimos a solução computadorizada de problemas de regressão, mostrando a saída de dados do Minitab relativa ao problema dos restaurantes Armand's Pizza Parlors. Neste apêndice, descrevemos as etapas necessárias para gerar a solução computadorizada com o Minitab. Primeiramente, os dados devem ser inseridos em uma planilha do Minitab. Os dados da população estudantil são inseridos na coluna C1 e os dados das vendas trimestrais são inseridos na coluna C2. Os nomes das variáveis Pop (População) e Sales (Vendas) são inseridos como títulos de coluna na planilha. Nas etapas subseqüentes, referimo-nos aos dados usando os nomes das variáveis Pop e Sales ou os indicadores de coluna C1 e C2. As etapas a seguir descrevem como usar o Minitab para produzir os resultados de regressão mostrados na Figura 12.10.

Etapa 1. Selecione o menu **Stat**
Etapa 2. Selecione o menu **Regression**
Etapa 3. Escolha a opção **Regression**
Etapa 4. Quando a caixa de diálogo Regression aparecer:
Digite Sales na caixa **Response**
Digite Pop na caixa **Predictors**
Dê um clique no botão **Options**

Quando a caixa de diálogo Regression-Options aparecer:

Digite 10 na caixa **Prediction intervals for new observations**
Dê um clique em **OK**

Quando a caixa de diálogo Regression reaparecer:

Dê um clique em **OK**

A caixa de diálogo Regression do Minitab oferece capacidades adicionais que podem ser obtidas selecionando-se as opções desejadas. Por exemplo, para obter uma plotagem residual que mostra o valor previsto da variável dependente $\hat{y}$ no eixo horizontal e os valores residuais no eixo vertical, a etapa 4 seria feita da seguinte maneira:

Etapa 4: Quando a caixa de diálogo Regression aparecer:
Digite Sales na caixa **Response**

Digite Pop na caixa **Predictors**
Dê um clique no botão **Graphs**

Quando a caixa de diálogo Regression-Graphs aparecer:

Selecione **Regular** sob a opção Residuals for Plots
Selecione **Residuals versus fits** sob a opção Residual Plots
Dê um clique em **OK**

Quando a caixa de diálogo Regression reaparecer:

Dê um clique em **OK**

Apêndice 12.2 – Análise de Regressão com o Excel

ARQUIVO DA INTERNET
Armand's

Neste apêndice, ilustraremos como a ferramenta Regressão do Excel pode ser usada para realizar os cálculos de análise de regressão do problema dos restaurantes Armand's Pizza Parlors. Consulte a Figura 12.14 à medida que descrevermos as etapas envolvidas. Os rótulos Restaurante, População e Vendas são inseridos nas células A1:C1 da planilha. Para identificar cada uma das dez observações, inserimos os números 1 a 10 nas células A2:A11.

Figura 12.14 Solução do Excel para o problema dos restaurantes Armand's Pizza Parlors

	A	B	C	D	E	F	G	H	I	J
1	**Restaurante**	**População**	**Vendas**							
2	1	2	58							
3	2	6	105							
4	3	8	88							
5	4	8	118							
6	5	12	117							
7	6	16	137							
8	7	20	157							
9	8	20	169							
10	9	22	149							
11	10	26	202							
12										
13	RESUMO DA SAÍDA									
14										
15	*Estatística de Regressão*									
16	R-Múltipla	0.9501								
17	R-Quadrado	0.9027								
18	R-Quadrado Ajustado	0.8906								
19	Erro Padrão	13.8293								
20	Observações	10								
21										
22	ANOVA									
23		*gl*	*SQ*	*MQ*	*F*	*Significância F*				
24	Regressão	1	14200	14200	74.2484	2.55E-05				
25	Resíduo	8	1530	191.25						
26	Total	9	15730							
27										
28		*Coeficientes*	*Erro Padrão*	*Estat. t*	*Valor P*	*95% Mín.*	*Máx. 95%*	*99.0% Mín.*	*99.0% Máx.*	
29	Intercepção	60	9.2260	6.5033	0.0002	38.7247	81.2753	29.0431	90.9569	
30	População	5	0.5803	8.6167	2.55E-05	3.6619	6.3381	3.0530	6.9470	
31										

Os dados amostrais são inseridos nas células B2:C11. As etapas a seguir descrevem como usar o Excel para produzir os resultados de regressão.

Etapa 1. Selecione o menu **Ferramentas**
Etapa 2. Escolha a opção **Análise de Dados**
Etapa 3. Escolha **Regressão** na lista de Ferramentas de Análise
Etapa 4. Dê um clique em **OK**
Etapa 5. Quando a caixa de diálogo Regressão aparecer:
Digite C1:C11 na caixa **Intervalo Y de Entrada**
Digite B1:B11 na caixa **Intervalo X de Entrada**
Selecione **Rótulos**
Selecione **Nível de Confiança**
Digite 99 na caixa **Nível de Confiança**
Selecione **Intervalo de Saída**
Digite A13 na caixa **Intervalo de Saída**
(Qualquer célula do canto superior esquerdo que indique onde a saída deve ser iniciada pode ser inserida aqui.)
Dê um clique em **OK**

A primeira seção da saída de dados, intitulada *Estatísticas da Regressão*, apresenta um resumo estatístico, por exemplo, o coeficiente de determinação (R-Quadrado). A segunda seção da saída, intitulada ANOVA, contém a tabela de análise de variância. A última seção da saída, a qual não tem um título, contém o coeficiente de regressão estimado e as informações correspondentes. Iniciaremos nossa discussão da interpretação da saída de regressão com a informação contida nas células A28:I30.

Interpretação da Saída de Dados da Equação de Regressão Estimada

O ponto em que a reta de regressão estimada intercepta o eixo y, $b_0 = 60$, é mostrado na célula B29, e a inclinação da reta de regressão estimada, $b_1 = 5$ é mostrada na célula B30. O rótulo Intercepto na célula A29 e o rótulo População na célula A30 são usados para identificar esses dois valores.

Na Seção 12.5, mostramos que o desvio padrão estimado de b_1 é $s_{b_1} = 0{,}5803$. Observe que o valor na célula C30 é 0,5803. O rótulo Erro Padrão na célula C28 é a maneira de o Excel indicar que o valor na célula C30 é o erro padrão, ou desvio padrão, de b_1. Lembre-se de que o teste t de uma relação significativa exigia o cálculo da estatística t, ou seja, $t = b_1/s_{b_1}$. Em relação aos dados dos restaurantes Armand's, o valor de t que calculamos foi $t = 5/0{,}5803 = 8{,}62$. O rótulo na célula D28, *Estatística t*, lembra-nos de que a célula D30 contém o valor da estatística t.

O valor na célula E30 é o valor associado ao teste t de significância. O Excel exibiu o valor p na célula E30 usando notação científica. Para obter o valor decimal, deslocamos a vírgula cinco casas decimais à esquerda, obtendo o valor 0,0000255. Uma vez que o valor $p = 0{,}0000255 < \alpha = 0{,}01$, podemos rejeitar H_0 e concluir que temos uma relação significativa entre a população de estudantes e as vendas trimestrais.

A informação nas células F28:I30 pode ser usada para desenvolver estimações por intervalo de confiança da intercepção com o eixo y e a inclinação da equação de regressão estimada. O Excel sempre apresenta os limites mínimo e máximo de um intervalo de confiança de 95%. Lembre-se de que na etapa 4, selecionamos Nível de Confiança e inserimos 99 na caixa Nível de Confiança. Em conseqüência, a ferramenta Regressão do Excel também fornece os limites mínimo e máximo de um intervalo de confiança de 99%. O valor na célula H30 é o limite mínimo da estimação por intervalo de confiança de 99% de β_1, e o valor na célula I30 é o limite máximo. Desse modo, após o arredondamento, a estimação por intervalo de confiança de 99% de β_1 varia de 3,05 a 6,95. Os valores nas células F30 e G30 fornecem os limites mínimo e máximo do intervalo de confiança de 95%. Assim, o intervalo de confiança de 95% varia de 3,66 a 6,34.

Interpretação da Saída de Dados ANOVA

A informação nas células A22:F26 é um resumo dos cálculos da análise de variância. As três fontes de variação são rotuladas de Regressão, Resíduo e Total. O rótulo *gl* na célula B23 refere-se a "graus de liberdade", o rótulo, o rótulo *SQ* na célula C23 corresponde à soma dos quadrados e o rótulo *MQ* na célula D23 refere-se à média quadrática.

O rótulo Significância *F* pode ser mais significativo se você imaginar o valor contido na célula F24 como o nível de significância observado para o teste *F*.

Na Seção 12.5, afirmamos que o erro médio quadrático, obtido ao dividir-se a soma dos quadrados dos erros ou resíduos por seus graus de liberdade, fornece uma estimativa de σ^2. O valor na célula D25, 191,25, é o erro médio quadrático da saída de dados da regressão correspondente aos restaurantes Armand's. Na Seção 12.5, mostramos que um teste F também poderia ser usado para testar a significância em uma regressão. O valor na célula F24, 0,0000255, é o valor p associado ao teste F de significância. Uma vez que o valor $p = 0{,}0000255 < \alpha = 0{,}01$, podemos rejeitar H_0 e concluir que temos uma relação significativa entre a população estudantil e as vendas trimestrais. O rótulo que o Excel usa para identificar o valor p, mostrado na célula F23, é *Significância F*.

Interpretação da Saída de Dados da Estatística de Regressão

O coeficiente de determinação, 0,9027, aparece na célula B:17; o rótulo correspondente, R-Quadrado, é indicado na célula A17. A raiz quadrada do coeficiente de determinação fornece o coeficiente de correlação amostral 0,9501, mostrado na célula B16. Observe que o Excel usa o rótulo R-Múltipla (célula A16) para identificar esse valor. Na célula A19, o rótulo Erro Padrão é usado para identificar o valor do erro padrão da estimativa exposta na célula B19. Desse modo, o desvio padrão da estimativa é 13,8293. Recomendamos ao leitor ter em mente que, na saída do Excel, o rótulo Erro Padrão aparece em dois lugares diferentes. Na Seção Estatística de Regressão da saída de dados, o rótulo Erro Padrão refere-se à estimativa σ. Na seção Equação de Regressão Estimada da saída de dados o rótulo *Erro Padrão* corresponde a s_{b_1}, o desvio padrão da distribuição amostral de b_1.

Regressão Múltipla

ESTATÍSTICA NA PRÁTICA

INTERNATIONAL PAPER*
Purchase, Nova York

A International Paper é a maior empresa mundial produtora de papel e de produtos florestais. A empresa emprega 117 mil pessoas em suas operações em aproximadamente 50 países e exporta seus produtos para mais de 130 nações. A International Paper produz materiais de construção – pranchas de madeira e madeira compensada, materiais de embalagem para bens de consumo –; copos e recipientes descartáveis; materiais de embalagem industrial – caixas de papel corrugado e contêineres de embarque –; bem como uma grande variedade de papéis para fotocopiadoras, impressoras, livros e materiais de propaganda.

Para fabricar produtos de papel, as usinas de polpa processam madeira picada e produtos químicos para produzir polpa de madeira (celulose). A celulose é então usada em uma usina de papel para produzir produtos de papel. Na produção de papel branco, a polpa deve ser branqueada para eliminar qualquer descoloração. Um agente fundamental no processo de branqueamento (*bleaching*) é o dióxido de cloro, o qual, em virtude de sua natureza combustível, geralmente é produzido em uma instalação da fábrica de celulose e depois é bombeado na forma de solução para uma estação de branqueamento. A fim de melhorar um dos processos usados na produção do dióxido de cloro, pesquisadores estudaram o controle e a eficiência do processo. Um dos aspectos do estudo examinou a taxa de suprimento de produtos químicos para a produção de dióxido de cloro.

Para produzir o dióxido de cloro, quatro produtos químicos fluem a taxas controladas para o gerador de dióxido de cloro. O dióxido de cloro produzido no gerador flui para um absorvedor, no qual água gelada absorve o gás dióxido de cloro para formar a solução de dióxido de cloro. A solução é então bombeada para

* Os autores agradecem a Marian Williams e a Bill Griggs por fornecer esta "Estatística na Prática". Essa aplicação foi desenvolvida originalmente na Champion International Corporation, a qual se tornou parte da International Paper em 2000.

a fábrica de papel. Uma parte fundamental no controle do processo envolve as taxas de suprimento de produtos químicos. Historicamente, operadores experientes definiam as taxas de suprimento dos produtos químicos, mas esse critério levou a um excesso de trabalho de controle da parte dos operadores. Conseqüentemente, os engenheiros químicos da usina solicitaram que um conjunto de equações de controle, uma para cada suprimento químico, fosse desenvolvido para auxiliar os operadores na tarefa de definir as taxas.

Usando análise de regressão múltipla, os analistas estatísticos desenvolveram uma equação de regressão múltipla estimada para cada um dos quatro produtos químicos empregados no processo. Cada equação relacionava a produção de dióxido de cloro com a quantidade de produto químico usado e o nível de concentração da solução de dióxido de cloro. O conjunto resultante de quatro equações foi programado em um computador em cada usina. No novo sistema, os operadores digitam a concentração da solução de dióxido de cloro e a taxa de produção desejadas; o software calcula então o suprimento de produto químico necessário para se obter a taxa de produção desejada. Depois que os operadores começaram a usar as equações de controle, a eficiência do gerador de dióxido de cloro se elevou, e o número de vezes que as concentrações permaneceram dentro de limites aceitáveis também se elevou significativamente.

Esse exemplo mostra como a análise de regressão múltipla pode ser usada para desenvolver um processo de branqueamento melhor para produzir produtos de papel branco. Neste capítulo, discutiremos como são usados softwares para essas finalidades. A maior parte dos conceitos apresentados no Capítulo 12, relativos à regressão linear simples, pode ser estendida diretamente ao caso das regressões múltiplas.

No Capítulo 12, apresentamos a regressão linear simples e demonstramos seu uso no desenvolvimento de uma equação de regressão estimada que descreva a relação entre duas variáveis. Lembre-se de que a variável prevista ou explicada é chamada variável dependente e a variável usada para prever ou explicar a variável dependente é denominada variável independente. Neste capítulo, prosseguimos nosso estudo da análise de regressão ao considerarmos situações que envolvem duas ou mais variáveis independentes. Essa matéria, intitulada **análise de regressão múltipla**, nos possibilita considerar mais fatores e, desse modo, obter melhores estimativas do que aquelas que são possíveis com a regressão linear simples.

13.1 MODELO DE REGRESSÃO MÚLTIPLA

Análise de regressão múltipla é o estudo de como a variável dependente y se relaciona com duas ou mais variáveis independentes. Em geral, usamos p para designar o número de variáveis independentes.

Modelo de Regressão e Equação de Regressão

Os conceitos de modelo de regressão e equação de regressão apresentados no capítulo anterior são aplicáveis ao caso da regressão múltipla. A equação que descreve como a variável dependente y está relacionada com as variáveis independentes $x_1, x_2, \ldots, x_p$ e um termo de erro denomina-se **modelo de regressão múltipla.** Iniciamos com a suposição de que o modelo de regressão múltipla assume a seguinte forma:

MODELO DE REGRESSÃO MÚLTIPLA

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \epsilon \qquad \textbf{(13.1)}$$

No modelo de regressão múltipla, $\beta_0, \beta_1, b_2, \ldots, \beta_p$ são os parâmetros e ϵ (a letra grega epsílon) é uma variável aleatória. Um exame minucioso desse modelo revela que y é uma função linear de $x_1, x_2, \ldots, x_p$ (a parte $\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p$) mais um termo de erro ϵ. O termo de erro é responsável pela variabilidade em y que não pode ser explicada pelo efeito linear das variáveis independentes p.

Na Seção 13.4, discutiremos as suposições referentes ao modelo de regressão múltipla e ϵ. Uma das suposições é que a média, ou valor esperado, de ϵ é zero. Uma das conseqüências dessa suposição é que a média, ou valor esperado, de y, designado $E(y)$, é igual a $\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p$. A equação que descreve como o valor médio de y está relacionado a $x_1, x_2, \ldots, x_p$ denomina-se **equação de regressão múltipla**.

EQUAÇÃO DE REGRESSÃO MÚLTIPLA

$$E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p \quad (13.2)$$

Equação de Regressão Múltipla Estimada

Se os valores de $\beta_0, \beta_1, \beta_2 \ldots, \beta_p$ fossem conhecidos, a Equação 13.2 poderia ser usada para calcular o valor médio de y em relação a valores dados de $x_1, x_2, \ldots, x_p$. Infelizmente, esses valores de parâmetro geralmente não serão conhecidos, e devem ser estimados a partir de dados amostrais. Uma variável aleatória simples é usada para calcular as estatísticas amostrais $b_0, b_1, b_2, \ldots, b_p$ que são utilizadas como estimadores por ponto dos parâmetros $\beta_0, \beta_1, \beta_2, \ldots, \beta_p$. Essas estatísticas amostrais fornecem a seguinte **equação de regressão múltipla estimada.**

EQUAÇÃO DE REGRESSÃO MÚLTIPLA ESTIMADA

$$\hat{y} = b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_p x_p \quad (13.3)$$

em que

$b_0, b_1, b_2, \ldots, b_p$ são as estimativas de $\beta_0, \beta_1, \beta_2, \ldots, \beta_p$
$\hat{y}$ = o valor estimado da variável dependente.

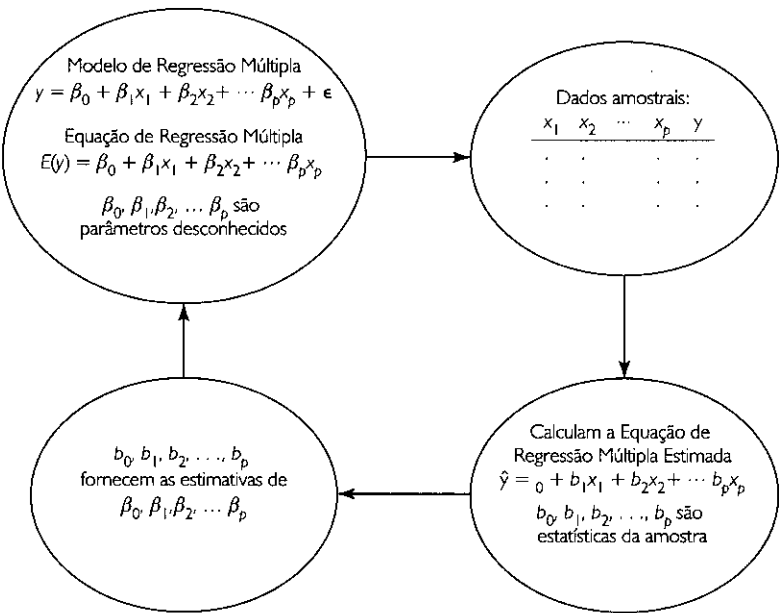
O processo de estimação para a regressão múltipla é mostrado na Figura 13.1.

13.2 MÉTODO DOS MÍNIMOS QUADRADOS

No Capítulo 12, usamos o **método dos mínimos quadrados** para desenvolver a equação de regressão estimada que mais bem aproximava a relação em linha reta entre as variáveis dependente e independente. Essa mesma abordagem é usada para desenvolver a equação de regressão múltipla estimada. O critério dos mínimos quadrados é reformulado da seguinte maneira:

CRITÉRIO DOS MÍNIMOS QUADRADOS

$$\text{mín } \Sigma(y_i - \hat{y}_i)^2 \quad (13.4)$$

Figura 13.1 O processo de estimação da regressão múltipla

Na regressão linear simples, b_0 e b_1 eram a estatística da amostra usada para estimar os parâmetros β_0 e β_1. A análise de regressão múltipla faz um paralelo com esse processo de inferência estatística, e $b_0, b_1, b_2, \ldots b_p$ denotam a estatística amostral utilizada para estimar os parâmetros $\beta_0, \beta_1, \beta_2, \ldots, \beta_p$.

em que

y_i = o valor observado da variável dependente para a i-ésima observação

$\hat{y}_i$ = o valor estimado da variável dependente para a i-ésima observação

Os valores estimados da variável dependente são calculados usando-se a equação de regressão múltipla estimada:

$$\hat{y} = b_0 + b_1x_1 + b_2x_2 + \cdots + b_px_p$$

Conforme mostra a Equação 13.4, o método dos mínimos quadrados usa dados amostrais para produzir os valores de $b_0, b_1, b_2, \ldots, b_p$ que transformam em um mínimo o somatório dos resíduos quadráticos [os desvios entre os valores observados da variável dependente (y_i) e os valores estimados da variável dependente ($\hat{y}_i$)].

No Capítulo 12, apresentamos fórmulas para calcular os estimadores por mínimos quadrados b_0 e b_1 para a equação de regressão linear simples estimada $\hat{y} = b_0 + b_1x$. Com relativamente poucos conjuntos de dados, fomos capazes de usar essas fórmulas para calcular b_0 e b_1 por meio de cálculos manuais. Na regressão múltipla, entretanto, a apresentação das fórmulas dos coeficientes de regressão $b_0, b_1, b_2, \ldots, b_p$ envolve o uso de álgebra matricial e está além do escopo deste livro. Portanto, ao apresentar a regressão múltipla, concentramo-nos em como é possível usar softwares para obter a equação de regressão estimada e outras informações. A ênfase será a maneira de interpretar a saída (*output*) de computador em vez de como efetuar os cálculos de regressão múltipla.

Exemplo: Butler Trucking Company

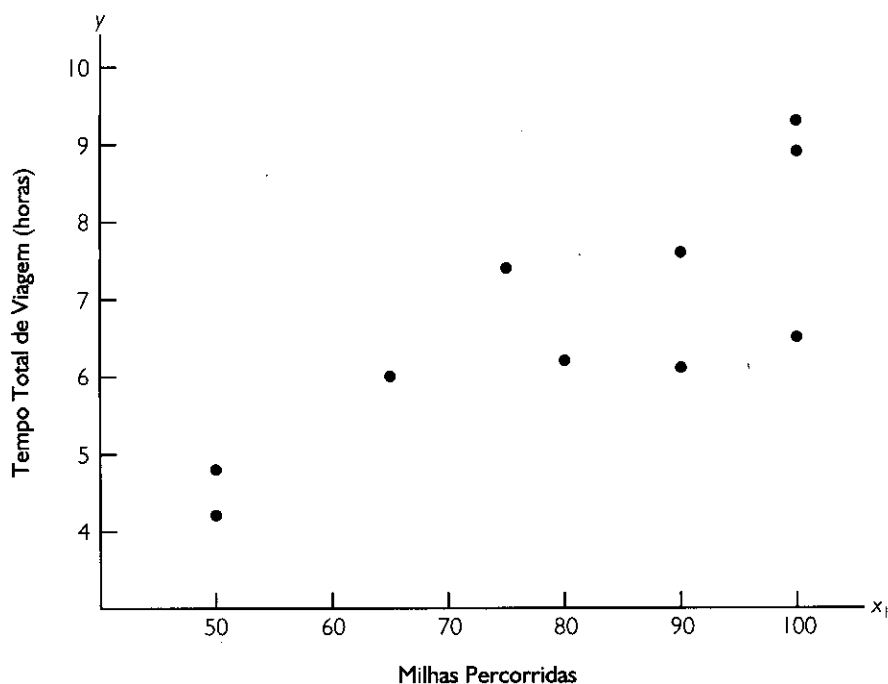
Como ilustração da análise de regressão múltipla, consideraremos um problema enfrentado pela Butler Trucking Company, uma empresa independente de transporte rodoviário de carga do sul da Califórnia. Uma parte importante dos negócios da Butler envolve entregas em toda a sua região. Para desenvolver melhores programas de trabalho, os gerentes querem estimar o tempo total diário das viagens de seus motoristas.

A princípio, os gerentes acreditavam que o tempo total diário das viagens estaria estreitamente relacionado com o número de milhas percorridas ao fazerem as entregas diárias. Uma amostra aleatória simples de dez tarefas de entrega forneceu os dados apresentados na Tabela 13.1 e o diagrama de dispersão da Figura 13.2. Depois de revisar esse diagrama de dispersão, os gerentes aventaram a hipótese de que o modelo de regressão linear simples $y = \beta_0 + \beta_1x_1 + e$ poderia ser usado para descrever a relação entre o tempo total de viagem (y) e o número de milhas percorridas (x_1).

ARQUIVO DA INTERNET
Butler

Tabela 13.1 Dados preliminares da Butler Trucking

Tarefa de Entrega	x_1 = Milhas Percorridas	y = Tempo de Viagem (horas)
1	100	9,3
2	50	4,8
3	100	8,9
4	100	6,5
5	50	4,2
6	80	6,2
7	75	7,4
8	65	6,0
9	90	7,6
10	90	6,1

Figura 13.2 Diagrama de dispersão dos dados preliminares da Butler Trucking

A fim de estimar os parâmetros β_0 e β_1, foi usado o método dos mínimos quadrados para desenvolver a equação de regressão estimada:

$$\hat{y} = b_0 + b_1x_1 \quad \textbf{(13.5)}$$

Na Figura 13.3, apresentamos o resultado do Minitab relativo à aplicação de regressão linear simples aos dados da Tabela 13.1. A equação de regressão estimada é:

$$\hat{y} = 1{,}27 + 0{,}0678x_1$$

No nível de significância 0,05, o valor F de 15,81 e seu correspondente valor p de 0,004 indicam que a relação é significativa; ou seja, podemos rejeitar H_0: $\beta_1 = 0$ porque o valor p é menor que $\alpha = 0{,}05$. Note que a mesma conclusão é obtida do valor t igual a 3,98 e seu valor p associado de 0,004. Desse modo, podemos concluir que a relação entre o tempo total de viagem e o número de milhas percorridas é significativa; tempos de viagem mais longos estão associados a mais milhas percorridas. Com um coeficiente de determinação (expresso como uma porcentagem) de R-sq = 66,4%, vemos que 66,4% da variabilidade relativa ao tempo de viagem podem ser explicados pelo efeito linear do número de milhas percorridas. Essa conclusão é razoavelmente boa, mas os gerentes quiseram considerar o acréscimo de uma segunda variável independente para explicar uma parte da variabilidade restante na variável dependente.

Ao tentar identificar outra variável independente, os gerentes acharam que o número de entregas também poderia contribuir para o tempo total de viagem. Os dados da Butler Trucking, com o acréscimo do número de entregas, são apresentados na Tabela 13.2. A solução computadorizada do Minitab, tendo as milhas percorridas (x_1) e o número de entregas (x_2) como variáveis independentes, é mostrada na Figura 13.4. A equação de regressão estimada é:

$$\hat{y} = -0{,}869 + 0{,}0611x_1 + 0{,}923x_2 \quad \textbf{(13.6)}$$

Na saída de dados do Minitab, os nomes das variáveis Miles e Time foram inseridos como cabeçalhos de coluna na planilha; desse modo, x_1 = Miles e y = Time.

Figura 13.3 Saída de dados do Minitab referente à Butler Trucking com uma variável independente

```
The regression equation is
Time = 1.27 + 0.0678 Miles

Predictor      Coef    SE Coef      T      p
Constant      1.274      1.401   0.91  0.390
Miles       0.06783    0.01706   3.98  0.004

S = 1.002   R-sq = 66.4%   R-sq(adj) = 62.2%

Analysis of Variance

SOURCE          DF       SS       MS      F      p
Regression       1   15.871   15.871  15.81  0.004
Residual Error   8    8.029    1.004
Total            9   23.900
```

Na próxima seção, discutiremos o uso do coeficiente de determinação múltiplo para medir o grau de eficiência de ajuste que é proporcionado pela equação de regressão estimada. Antes de fazê-lo, vamos examinar mais cuidadosamente os valores de b_1 = 0,0611 e b_2 = 0,923 na Equação 13.6.

Nota sobre a Interpretação de Coeficientes

A esta altura, pode-se fazer uma observação sobre a relação entre a equação de regressão estimada, tendo somente as milhas percorridas como a variável independente, e a equação que inclui o número de entregas como a segunda variável independente. O valor de b_1 não é o mesmo em ambos os casos. Na regressão linear simples, interpretamos b_1 como uma estimativa da alteração em y correspondente à alteração de uma unidade na variável independente. Na análise de regressão múltipla, a interpretação deve ser bastante modificada. Ou seja, na análise de regressão múltipla, interpretamos cada coeficiente de regressão da seguinte maneira; b_1 representa uma estimativa da alteração em y correspondente à alteração de uma unidade em x_i quando todas as outras variáveis independentes se mantêm constantes. No exemplo da Butler Trucking envolvendo duas variáveis independentes, b_1 é igual a 0,0611.

ARQUIVO DA INTERNET
Butler

Tabela 13.2 Dados da Butler Trucking tendo as milhas percorridas (x_1) e o número de entregas (x_2) como variáveis independentes

Tarefa de Entrega	x_1 = Milhas Percorridas	x_2 = Número de Entregas *Deliveries*	y = Tempo de Viagem (horas)
1	100	4	9,3
2	50	3	4,8
3	100	4	8,9
4	100	2	6,5
5	50	2	4,2
6	80	2	6,2
7	75	3	7,4
8	65	4	6,0
9	90	3	7,6
10	90	2	6,1

Figura 13.4 Saída de dados do Minitab para a Butler Trucking com duas variáveis independentes

```
The regression equation is
Time = - 0.869 + 0.0611 Miles + 0.923 Deliveries

Predictor        Coef    SE Coef       T       p
Constant      -0.8687     0.9515   -0.91   0.392
Miles        0.061135   0.009888    6.18   0.000
Deliveries     0.9234     0.2211    4.18   0.004

S = 0.5731    R-sq = 90.4%    R-sq(adj) = 87.6%

Analysis of Variance

SOURCE          DF       SS        MS        F       p
Regression       2   21.601    10.800    32.88   0.000
Residual Error   7    2.299     0.328
Total            9   23.900
```

Na saída de dados do Minitab os nomes das variáveis Miles e Time foram inseridas como cabeçalhos de coluna na planilha; desse modo, x_1 = Miles, x_2 = Deliv e y = Time.

Assim, 0,0611 é a estimativa do aumento esperado no tempo de viagem correspondente ao aumento de 1 milha na distância percorrida quando o número de entregas é mantido constante. Similarmente, desde que b_2 é igual a 0,923, a estimativa do tempo de viagem esperado correspondente ao aumento de uma entrega quando o número de milhas percorridas é mantido constante é 0,923 horas.

Exercícios

Nota para o estudante: Os exercícios que envolvem dados, nesta e nas seções subseqüentes, foram idealizados para serem resolvidos com o auxílio de um software de computador.

Métodos

1. A equação de regressão estimada de um modelo que envolve duas variáveis independentes e dez observações é a seguinte:

$$\hat{y} = 29{,}1270 + 0{,}5906x_1 + 0{,}4980x_2$$

a. Interprete b_1 e b_2 nessa equação de regressão estimada.
b. Estime y quando $x_1 = 180$ e $x_2 = 310$.

2. Considere os seguintes dados referentes a uma variável dependente y e duas variáveis independentes, x_1 e x_2:

AUTOTESTE

ARQUIVO DA INTERNET
Exer2

x_1	x_2	y
30	12	94
47	10	108
25	17	112
51	16	178
40	5	94
51	19	175
74	7	170
36	12	117
59	13	142
76	16	211

a. Desenvolva uma equação de regressão estimada relacionando y com x_1. Estime y se $x_1 = 45$.
b. Desenvolva uma equação de regressão estimada relacionando y com x_2. Estime y se $x_2 = 15$.
c. Desenvolva uma equação de regressão estimada relacionando y com x_1 e com x_2. Estime y se $x_1 = 45$ e $x_2 = 15$.

3. Em uma análise de regressão envolvendo 30 observações foi obtida a seguinte equação de regressão estimada.

$$\hat{y} = 17{,}6 + 3{,}8x_1 + 2{,}3x_2 + 7{,}6x_3 + 2{,}7x_4$$

a. Interprete b_1, b_2, b_3 e b_4 nessa equação de regressão estimada.
b. Estime y quando $x_1 = 10$, $x_2 = -5$, $x_3 = 1$ e $x_4 = 2$.

Aplicações

4. Uma loja de calçados desenvolveu a seguinte equação de regressão estimada relacionando as vendas com o investimento em estoques e os gastos de propaganda:

$$\hat{y} = 25 + 10x_1 + 8x_2$$

em que

x_1 = investimento em estoques (US$ 1.000)
x_2 = gastos de propaganda (US$ 1.000)
y = vendas (US$ 1.000)

a. Estime as vendas resultantes de um investimento de US$ 15 mil em estoques e um orçamento de propaganda de US$ 10 mil.
b. Interprete b_1 e b_2 nessa equação de regressão estimada.

AUTOTESTE

5. O proprietário da Showtime Movie Theaters, Inc., gostaria de estimar semanalmente a receita bruta em função dos gastos de propaganda. Os dados históricos de uma amostra de oito semanas são os seguintes:

ARQUIVO DA INTERNET
Showtime

Receita Bruta Semanal (em milhares de dólares)	Propaganda de Televisão (em milhares de dólares)	Propaganda de Jornal (em milhares de dólares)
96	5,0	1,5
90	2,0	2,0
95	4,0	1,5
92	2,5	2,5
95	3,0	3,3
94	3,5	2,3
94	2,5	4,2
94	3,0	2,5

a. Desenvolva uma equação de regressão estimada, sendo a quantia gasta em propaganda de televisão a variável independente.
b. Estabeleça uma equação de regressão estimada, sendo a quantia gasta em propaganda de televisão e a quantia gasta em propaganda de jornal as variáveis independentes.
c. O coeficiente da equação de regressão estimada correspondente aos gastos de propaganda de televisão é idêntico nos itens (a) e (b)? Interprete o coeficiente em cada caso.
d. Qual é a estimativa da receita bruta semanal de uma semana em que são gastos US$ 3.500 em propaganda de televisão e US$ 1.800 em propaganda de jornal?

6. No beisebol, o sucesso de uma equipe freqüentemente é considerado uma função do seu desempenho para rebater e arremessar a bola. Uma medida do desempenho da equipe rebatedora é o número de *home runs*[1] que esse time faz, e uma medida do desempenho da equipe arremessadora (*pitchers*) é a média de *runs* conquistados por ela. Geralmente, acredita-se que as equipes que acertam mais *home runs* e que recebem uma média menor de *runs* da outra equipe vencerão uma porcentagem maior dos jogos disputados. Os dados a seguir apresentam a porcentagem de partidas ganhas (PPG), o número de *home runs* (HR) feitos pela equipe e a média de *runs* recebidos (MRR) da equipe adversária referentes às 16 equipes da National League na temporada de 2003 da Major League Baseball (http://www.usatoday, 7 de janeiro de 2004).

[1] NT: *Home run* – Uma rebatida certeira, para longe, que permite ao rebatedor percorrer todas as bases e marcar um *run* [pontuação por percorrer (tocar) de maneira bem-sucedida todas as bases] (Beisebol).

Equipe	PPG (% de Partidas Ganhas)	HR (Número de *Home Runs* Feitos)	MRR (Média de *Runs* Recebidos da Equipe Adversária)	Equipe	PPG (% de Partidas Ganhas)	HR (Número de *Home Runs* Feitos)	MRR (Média de *Runs* Recebidos da Equipe Adversária)
Arizona	0,519	152	3,857	Milwaukee	0,420	196	5,058
Atlanta	0,623	235	4,106	Montreal	0,512	144	4,027
Chicago	0,543	172	3,842	Nova York	0,410	124	4,517
Cincinnati	0,426	182	5,127	Philadelphia	0,531	166	4,072
Colorado	0,457	198	5,269	Pittsburgh	0,463	163	4,664
Florida	0,562	157	4,059	San Diego	0,395	128	4,904
Houston	0,537	191	3,880	San Francisco	0,621	180	3,734
Los Angeles	0,525	124	3,162	St. Louis	0,525	196	4,642

ARQUIVO DA INTERNET
MLB

a. Determine a equação de regressão estimada que possa ser usada para prever a porcentagem de partidas ganhas, dado o número de *home runs* da equipe.
b. Estabeleça a equação de regressão estimada que possa ser utilizada para prever a porcentagem de partidas ganhas, dada a média de *runs* recebidos da equipe que faz os arremessos.
c. Elabore a equação de regressão estimada que possa ser usada para prever a porcentagem de partidas ganhas, dado o número de *home runs* e a média de *runs* recebidos da equipe que faz os arremessos.
d. Na temporada de 2003, San Diego venceu somente 39,5% das partidas que disputou, o índice mais baixo da National League. Para melhorar seu desempenho no ano seguinte, a equipe está tentando contratar novos jogadores que aumentem para 180 o número de *home runs* feitos pela equipe e diminuam para 4,0 a média de *runs* recebidos da equipe que faz os arremessos. Use a equação de regressão estimada desenvolvida no item (c) para estimar a porcentagem de jogos que San Diego vencerá se a equipe fizer 180 *home runs* e se a média de *runs* recebidos da equipe adversária for 4,0.

7. Os desenhistas (*designers*) de mochilas *backpack* usam materiais exóticos, por exemplo, supernáilon Delrin, polietileno de alta densidade, alumínio de aviação e espuma termomoldada para produzir mochilas que se ajustam confortavelmente e distribuem o peso para eliminar pontos de pressão. Os dados a seguir apresentam a capacidade (centímetros cúbicos), avaliação do conforto e o preço de dez mochilas *backpack* testadas pela *Outside Magazine*. O conforto foi medido usando-se uma escala de 1 a 5, sendo que a classificação 1 representa um conforto médio e a classificação 5 representa um conforto excelente (*Outside Buyer's Guide*, 2001).

Fábrica e Modelo	Capacidade	Conforto	Preço (US$)
Camp Trails Paragon II	70.955	2	190
EMS 5500	90.128	3	219
Lowe Alpomayo 90 + 20	90.128	4	249
Marmot Muir	77.019	3	249
Kelly Bigfoot 5200	85.212	4	250
Gregory Whitney	90.128	4	340
Osprey 75	77.019	4	389
Arc'Teryx Bora 95	90.128	5	395
Dana Design Terraplane LTW	95.044	5	439
The Works @ Mystery Ranch Jazz	81.935	5	525

ARQUIVO DA INTERNET
Backpack

a. Determine a equação de regressão estimada que pode ser usada para prever o preço de uma mochila *backpack*, dada a capacidade e a avaliação do conforto.
b. Interprete b_1 e b_2.
c. Preveja o preço de uma mochila *backpack* com capacidade para 73.741 centímetros cúbicos, sendo 4 a avaliação do conforto.

8. A tabela seguinte apresenta o retorno anual, a avaliação da segurança (0 = a mais arriscada, 10 = a mais segura) e a taxa de despesa anual referentes a 20 fundos de investimento estrangeiros (*Mutual Funds*, março de 2000).

ARQUIVO DA INTERNET
ForFunds

Fundo	Avaliação de Segurança	Taxa de Despesa Anual (%)	Retorno Anual (%)
Accessor Int'l Equity "Adv"	7,1	1,59	49
Aetna "I" International	7,2	1,35	52
Amer Century Int'l Discovery "Inv"	6,8	1,68	89
Columbia International Stock	7,1	1,56	58
Concert Inv "A" Int'l Equity	6,2	2,16	131
Dreyfus Founders Int'l Equity "F"	7,4	1,80	59
Driehaus International Growth	6,5	1,88	99
Excelsior "Inst" Int'l Equity	7,0	0,90	53
Julius Baer International Equity	6,9	1,79	77
Marshall International Stock "Y"	7,2	1,49	54
MassMutual Int'l Equity "S"	7,1	1,05	57
Morgan Grenfell Int'l Sm Cap "Inst"	7,7	1,25	61
New England "A" Int'l Equity	7,0	1,83	88
Pilgrim Int'l Small Cap "A"	7,0	1,94	122
Republic International Equity	7,2	1,09	71
Sit International Growth	6,9	1,50	51
Smith Barney "A" Int'l Equity	7,0	1,28	60
State St Research "S" Int'l Equity	7,1	1,65	50
Strong International Stock	6,5	1,61	93
Vontobel International Equity	7,0	1,50	47

a. Desenvolva uma equação de regressão estimada relacionando o retorno anual com a avaliação da segurança e a taxa de despesa anual.

b. Estime o retorno anual de uma firma que tem uma avaliação de segurança igual a 7,5 e taxa de despesa anual igual a 2.

9. Dois especialistas apresentaram listas subjetivas de distritos escolares que consideram estar entre os melhores do país. Em relação a cada distrito escolar, foram apresentados o tamanho médio da classe, a pontuação média no SAT[2] e a porcentagem de estudantes que freqüentavam um curso superior de quatro anos.

ARQUIVO DA INTERNET
Schools

Distrito	Tamanho Médio da Classe	Pontuação Média no SAT	% dos que Freqüentam um Curso Superior de Quatro Anos
Blue Springs, MO	25	1083	74
Garden City, NY	18	997	77
Indianapolis, IN	30	716	40
Newport Beach, CA	26	977	51
Novi, MI	20	980	53
Piedmont, CA	28	1042	75
Pittsburg, PA	21	983	66
Scarsdale, NY	20	1110	87
Wayne, PA	22	1040	85
Weston, MA	21	1031	89
Farmingdale, NY	22	947	81
Mamaroneck, NY	20	1000	69
Mayfield, OH	24	1003	48
Morristown, NJ	22	972	64
New Rochelle, NY	23	1039	55
Newtown Square, PA	17	963	79
Omaha, NE	23	1059	81
Shaker Heights, OH	23	940	82

[2] NT: SAT (Sigla de *Scholastic Aptitude Test*) – Um exame usado pelas universidades como parte do processo de seleção de estudantes para admissão ao curso superior.

a. Usando esses dados, desenvolva uma equação de regressão estimada relacionando a porcentagem de estudantes que freqüentam um curso superior de quatro anos com o tamanho médio da classe e a pontuação média no SAT.
b. Estime a porcentagem de estudantes que freqüentam um curso superior de quatro anos se o tamanho médio da classe é 20 e a pontuação média no SAT é 1.000.

10. A National Basketball Association (NBA) registra uma série de estatísticas a respeito de cada time. Quatro dessas estatísticas são a porcentagem de jogos ganhos (PJG), a porcentagem de "*field goals*[3] (FG%), a porcentagem de lances de três pontos feitos pelo time adversário (%3Pt Adv) e o número de *turnovers*[4] cometidos pelo time adversário (Turnover Adv). Os dados a seguir apresentam os valores dessas estatísticas correspondentes às 29 equipes da NBA e se referem a uma parte da temporada de 2004 (http://www.nba.com, 3 de janeiro de 2004).

Equipe	PJG	FG%	% 3Pt Adv	*Turnover* Adv	Equipe	PJG	FG%	% 3Pt Adv	*Turnover* Adv
Atlanta	0,265	0,435	0,346	13,206	Minnesota	0,677	0,473	0,348	13,839
Boston	0,471	0,449	0,369	16,176	New Jersey	0,563	0,435	0,338	17,063
Chicago	0,313	0,417	0,372	15,031	New Orleans	0,636	0,421	0,330	16,909
Cleveland	0,303	0,438	0,345	12,515	Nova York	0,412	0,442	0,330	13,588
Dallas	0,581	0,439	0,332	15,000	Orlando	0,242	0,417	0,360	14,242
Denver	0,606	0,431	0,366	17,818	Philadelphia	0,438	0,428	0,364	16,938
Detroit	0,606	0,423	0,262	15,788	Phoenix	0,364	0,438	0,326	16,515
Golden State	0,452	0,445	0,384	14,290	Portland	0,484	0,447	0,367	12,548
Houston	0,548	0,426	0,324	13,161	Sacramento	0,724	0,466	0,327	15,207
Indiana	0,706	0,428	0,317	15,647	San Antonio	0,688	0,429	0,293	15,344
L.A. Clippers	0,464	0,424	0,326	14,357	Seattle	0,533	0,436	0,350	16,767
L.A. Lakers	0,724	0,465	0,323	16,000	Toronto	0,516	0,424	0,314	14,129
Memphis	0,485	0,432	0,358	17,848	Utah	0,531	0,456	0,368	15,469
Miami	0,424	0,410	0,369	14,970	Washington	0,300	0,411	0,341	16,133
Milwaukee	0,500	0,438	0,349	14,750					

ARQUIVO DA INTERNET NBA

Legenda: PJG – Porcentagem de Jogos Ganhos
FG% – Porcentagem de *Field Goals*
%3Pt Adv – Porcentagem de lances de três pontos feitos pelo time adversário
Turnover Adv – Número de *turnovers* cometidos pelo time adversário

a. Determine a equação de regressão estimada que possa ser usada para prever a porcentagem de jogos ganhos, dada a porcentagem de *field goals* feitos pelo time.
b. Forneça uma interpretação da inclinação da equação de regressão estimada desenvolvida no item (a).
c. Estipule a equação de regressão estimada que possa ser usada para prever a porcentagem de jogos ganhos, dada a porcentagem de *field goals* feitos pela equipe, a porcentagem de lances de três pontos feitos pelo adversário da equipe, e o número de *turnovers* cometidos pelo adversário da equipe.
d. Discuta as implicações práticas da equação de regressão estimada desenvolvida no item (c).
e. Estime a porcentagem de jogos ganhos pela equipe, com os seguintes valores para as três variáveis independentes: FG% = 0,45, %3Pt Adv = 0,34 e *Turnover* Adv = 17.

13.3 COEFICIENTE DE DETERMINAÇÃO MÚLTIPLO

Na regressão linear simples, mostramos que a soma total dos quadrados pode ser dividida em dois componentes: a soma dos quadrados da regressão e a soma dos quadrados dos erros. O mesmo procedimento se aplica à soma dos quadrados na regressão múltipla.

[3] NT: *Field goal* – Um lance feito da quadra que vale dois pontos, e se for de certa distância (no basquetebol profissional, no mínimo 7,62 m), três pontos (Basquete).
[4] NT: *Turnover* Bola perdida – O time perde a posse de bola devido a uma falha ou falta cometida (Basquete).

RELAÇÃO ENTRE SST, SSR E SSE

$$SST = SSR + SSE \quad (13.7)$$

em que

$$SST = \text{soma total dos quadrados} = \Sigma(y_i - \bar{y})^2$$
$$SSR = \text{soma dos quadrados da regressão} = \Sigma(\hat{y}_i - \bar{y})^2$$
$$SSE = \text{soma dos quadrados dos erros} = \Sigma(y_i - \hat{y}_i)^2$$

Em virtude da dificuldade dos cálculos das três somas de quadrados, recorremos a softwares para determinar esses valores. A parte da análise de variância da saída de dados do Minitab da Figura 13.4 apresenta os três valores do problema da Butler Trucking com duas variáveis independentes: SST = 23,900, SSR = 21,601 e SSE = 2,299. Com somente uma variável independente (o número de milhas percorridas), a saída do Minitab da Figura 13.3 mostra que SST = 23,900, SSR = 15,891 e SSE = 8,029. O valor de SST é idêntico em ambos os casos porque ele não depende de $\hat{y}$, mas SSR aumenta e SSE decresce quando uma segunda variável (número de entregas) é acrescentada. A implicação é que a equação de regressão múltipla estimada proporciona melhor ajuste para os dados observados.

No Capítulo 14, usamos o coeficiente de determinação r^2 = SSR/SST, para medir a eficiência de ajuste da equação de regressão estimada. O mesmo conceito se aplica à regressão múltipla. O termo **coeficiente de determinação múltiplo** indica que estamos medindo a eficiência de ajuste da equação de regressão múltipla estimada. O coeficiente de determinação múltiplo, designado R^2, é calculado da seguinte maneira:

COEFICIENTE DE DETERMINAÇÃO MÚLTIPLO

$$R^2 = \frac{SSR}{SST} \quad (13.8)$$

O acréscimo de variáveis independentes faz que os erros de previsão se tornem menores, reduzindo assim a soma de quadrados dos erros, SSE. Uma vez que SSR = SST – SSE, quando SSE toma-se menor, SSR toma-se maior, fazendo com que R^2 = SSR/SST se eleve.

O coeficiente de determinação múltiplo pode ser interpretado como a proporção da variabilidade da variável dependente que pode ser explicada pela equação de regressão múltipla estimada. Portanto, quando é multiplicado por 100, ele pode ser interpretado como a porcentagem da variabilidade em y que pode ser explicada pela equação de regressão estimada.

No exemplo da Butler Trucking, com duas variáveis independentes, sendo SSR = 21,601 e SST = 23,900, temos:

$$R^2 = \frac{21{,}601}{23{,}900} = 0{,}904$$

Portanto, 90,4% da variabilidade no tempo de viagem y são explicados pela equação de regressão múltipla estimada, sendo as milhas percorridas e o número de entregas as variáveis independentes. Na Figura 13.4, notamos que o coeficiente de determinação múltiplo também é fornecido pela saída do Minitab; ele é designado por R-sq = 90,4%.

Se uma variável é acrescentada ao modelo, R^2 toma-se maior mesmo que a variável acrescentada não seja estatisticamente significativa. O coeficiente de determinação múltiplo ajustado compensa o número de variáveis independentes no modelo.

A Figura 13.3 mostra que o valor de R-sq da equação de regressão estimada com somente uma variável independente, isto é, o número de milhas percorridas (x_1), é 66,4%. Desse modo, a porcentagem de variabilidade nos tempos de viagem que é explicada pela equação de regressão estimada se eleva de 66,4% para 90,4% quando o número de entregas é adicionado como uma segunda variável independente. Em geral, R^2 sempre se eleva quando são adicionadas variáveis independentes ao modelo.

Muitos analistas preferem ajustar R^2 ao número de variáveis independentes a fim de evitar uma superestimação do impacto de se adicionar uma variável independente à quantidade de variabilidade explicada pela equação de regressão estimada. Com n denotando o número de variações e p denotando o número de variáveis independentes, o **coeficiente de determinação múltiplo ajustado** é calculado da seguinte maneira:

COEFICIENTE DE DETERMINAÇÃO MÚLTIPLO AJUSTADO

$$R_a^2 = 1 - (1 - R^2)\frac{n-1}{n-p-1} \quad (13.9)$$

Quanto ao exemplo da Butler Trucking, com $n = 10$ e $p = 2$, temos:

$$R_a^2 = 1 - (1 - 0{,}904)\frac{10 - 1}{10 - 2 - 1} = 0{,}88$$

Assim, depois de ajustarmos as duas variáveis independentes, obtemos um coeficiente de determinação múltiplo ajustado igual a 0,88. Esse valor é fornecido pela saída do Minitab da Figura 13.4 como R-sq(adj) = 87,6%; O valor que calculamos difere porque usamos um valor arredondado de R^2 no cálculo.

NOTAS E COMENTÁRIOS

Se o valor de R^2 for pequeno e o modelo contiver um número grande de variáveis independentes, o coeficiente de determinação ajustado pode assumir um valor negativo; nesses casos, o Minitab fixa o coeficiente de determinação ajustado em zero.

Exercícios

Métodos

11. No exercício 1 foi apresentada a equação de regressão estimada baseada em dez observações:

$$\hat{y} = 29{,}1270 + 0{,}5906x_1 + 0{,}4980x_2$$

Os valores de SST e SSR são 6724,125 e 6216,375, respectivamente.

a. Encontre a SSE.
b. Calcule R^2.
c. Calcule R_a^2.
d. Comente a eficiência de ajuste.

12. No exercício 2, foram fornecidas dez observações de uma variável dependente y e duas variáveis independentes x_1 e x_2; para esses dados, SST = 15.182,9 e SSR = 14.052,2.

AUTOTESTE

a. Calcule R^2.
b. Calcule R_a^2.
c. A equação de regressão estimada explica a grande quantidade de variabilidade dos dados? Explique.

13. No exercício 3, foi apresentada a seguinte equação de regressão estimada baseada em 30 observações:

$$\hat{y} = 17{,}6 + 3{,}8x_1 + 2{,}3x_2 + 7{,}6x_3 + 2{,}7x_4$$

Os valores de SST e SSR são 1.805 e 1.760, respectivamente.

a. Calcule R^2.
b. Calcule R_a^2.
c. Comente a eficiência de ajuste.

Aplicações

14. No exercício 4, foi apresentada a seguinte equação de regressão estimada relacionando as vendas com o investimento em estoques e os gastos de propaganda:

$$\hat{y} = 25 + 10x_1 + 8x_2$$

Os dados usados para desenvolver o modelo foram extraídos de uma pesquisa de dez lojas; em relação a esses dados, SST = 16.000 e SSR = 12.000.

a. Calcule R^2 em relação à equação de regressão estimada dada.
b. Calcule R_a^2.
c. O modelo parece explicar uma grande quantidade de variabilidade nos dados? Explique.

AUTOTESTE

15. No exercício 5, o proprietário da Showtime Movie Theaters, Inc., utilizou a análise de regressão múltipla para prever a receita bruta (y) em função da propaganda de televisão (x_1) e da propaganda em jornais (x_2). A equação de regressão estimada foi:

$$\hat{y} = 83{,}2 + 2{,}29x_1 + 1{,}30x_2$$

ARQUIVO DA INTERNET
Showtime

A solução computadorizada forneceu SST = 25,5 e SSR = 23,435.

a. Calcule e interprete R^2 e R^2_a.

b. Quando a propaganda de televisão era a única variável independente, $R^2 = 0{,}653$ e $R^2_a = 0{,}595$. Você prefere os resultados da regressão múltipla? Explique.

ARQUIVO DA INTERNET
MLB

16. No exercício 6, foram apresentados dados sobre a porcentagem de jogos ganhos, o número de *home runs* do time e a média de *runs* aplicados pela equipe arremessadora, correspondentes às 16 equipes da National League na temporada de beisebol da Major League para 2003 (http://www.usatoday, 7 de janeiro de 2004).

a. A equação de regressão estimada que utiliza somente o número de *home runs* como variável independente para prever a porcentagem de jogos ganhos proporcionou um bom ajuste? Explique.

b. Discuta os benefícios de se usar tanto o número de *home runs* efetuados como a média de *runs* recebidos da equipe adversária para prever a porcentagem de jogos ganhos.

ARQUIVO DA INTERNET
Schools

17. No exercício 9, foi desenvolvida uma equação de regressão estimada relacionando a porcentagem de estudantes que freqüentam um curso superior de quatro anos com o tamanho médio da classe e a pontuação média no SAT.

a. Calcule e interprete R^2 e R^2_a.

b. A equação de regressão estimada proporciona um bom ajuste para os dados? Explique.

ARQUIVO DA INTERNET
NBA

18. Consulte o exercício 10, no qual foram registrados dados sobre uma série de estatísticas correspondentes aos 29 times da National Basketball Association, relativas a uma parte da temporada de 2004 (http://www.nba.com, 3 de janeiro de 2004).

a. No item (c) do exercício 10, foi desenvolvida uma equação de regressão estimada relacionando a porcentagem de jogos ganhos dada a porcentagem de *field goals* feitos pela equipe, a porcentagem de lances de três pontos feitos pela equipe adversária e o número de *turnovers* cometidos pela equipe adversária. Quais são os valores de R^2 e R^2_a?

b. A equação de regressão estimada proporciona um bom ajuste aos dados? Explique.

13.4 SUPOSIÇÕES DO MODELO

Na Seção 13.1, apresentamos o seguinte modelo de regressão múltipla:

MODELO DE REGRESSÃO MÚLTIPLA

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \epsilon \qquad \textbf{(13.10)}$$

As suposições sobre o termo de erro e no modelo de regressão múltipla fazem um paralelo com as do modelo de regressão linear simples.

SUPOSIÇÕES SOBRE O TERMO DE ERRO E NO MODELO DE REGRESSÃO MÚLTIPLA $y = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p + \epsilon$

1. O erro ϵ é uma variável aleatória com média, ou valor esperado, igual a zero; ou seja $E(\epsilon) = 0$.
Implicação: Para dados valores de $x_1, x_2, \ldots, x_p$, o valor esperado, ou média, de y é dado por

$$E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p \qquad \textbf{(13.11)}$$

A Equação 13.11 é a equação de regressão múltipla que apresentamos na Seção 13.1. Nessa equação, $E(y)$ representa a média de todos os valores possíveis de y que poderiam ocorrer para determinados valores de $x_1, x_2, \ldots, x_p$.

2. A variância de ϵ é designada σ^2 e é idêntica para todos os valores das variáveis independentes $x_1, x_2, \ldots, x_p$.
Implicação: A variância de y nas proximidades da linha de regressão é igual a σ^2 e é idêntica para todos os valores de $x_1, x_2, \ldots, x_p$.

3. Os valores de ϵ são independentes.
Implicação: O tamanho do erro de um conjunto de valores em particular das variáveis independentes não está relacionado com o tamanho do erro de qualquer outro conjunto de valores.

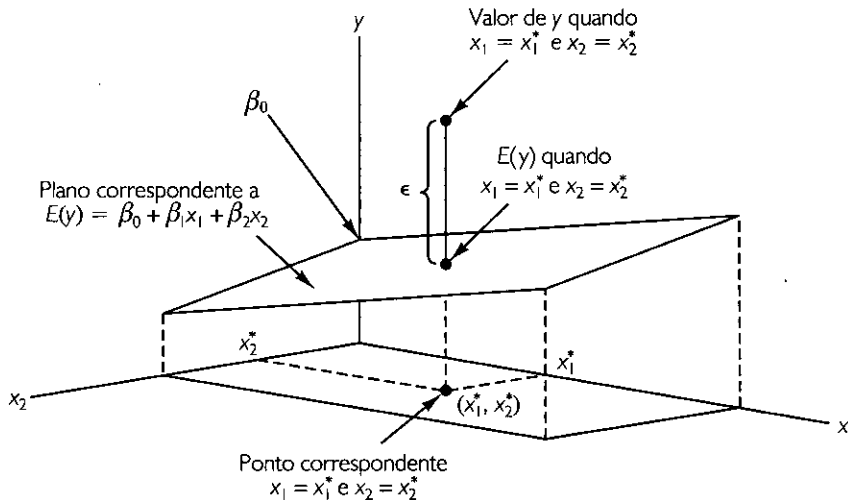
4. O erro ϵ é uma variável aleatória normalmente distribuída que reflete o desvio entre o valor y e o valor esperado de y dado por $\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p$.
Implicação: Uma vez que $\beta_0, \beta_1, \ldots, \beta_p$ são constantes para determinados valores de $x_1, x_2, \ldots, x_p$, a variável dependente y também é uma variável aleatória normalmente distribuída.

Para obter mais *insight* sobre a forma da relação dada pela Equação 13.11, considere a seguinte equação de regressão múltipla de duas variáveis independentes:

$$E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

O gráfico dessa equação é um plano em um espaço tridimensional. A Figura 13.5 apresenta um exemplo desse gráfico. Observe que o valor de ϵ mostrado é a diferença entre o valor de y real e o valor de y esperado, $E(y)$, quando $x_1 = x_1^*$ e $x_2 = x_2^*$.

Na análise de regressão, o termo *variável de resposta* freqüentemente é usado em lugar do termo *variável dependente*. Além disso, desde que a equação de regressão múltipla gere um plano ou superfície, seu gráfico se denomina *superfície de resposta*.

Figura 13.5 Gráfico da equação de regressão da análise de regressão múltipla com duas variáveis independentes

13.5 TESTE DE SIGNIFICÂNCIA

Nesta seção, mostramos como realizar testes de significância de uma relação de regressão múltipla. Os testes de significância que usamos na regressão linear simples foram um teste t e um teste F. Na regressão linear simples, ambos os testes produzem a mesma conclusão; ou seja, se a hipótese nula for rejeitada, concluiremos que $b_1 \neq 0$. Na análise de regressão múltipla, o teste t e o teste F têm propósitos diferentes.

1. O teste F é utilizado para determinar se existe uma relação significativa entre a variável dependente e o conjunto de todas as variáveis independentes; referirmo-nos ao teste F como teste de *significância global*.
2. Se o teste F exibir uma significância global, o teste t é usado para determinar se cada uma das variáveis independentes individuais é significativa. Um teste t separado é realizado para cada uma das variáveis independentes do modelo; referimo-nos a cada um desses testes t como teste de *significância individual*.

No material que apresentamos a seguir, explicaremos o teste F e o teste t e aplicaremos cada um ao exemplo da Butler Trucking Company.

Teste *F*

O modelo de regressão múltipla, de acordo com o que foi definido na Seção 13.4, é:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \epsilon$$

As hipóteses do teste *F* envolvem os parâmetros do modelo de regressão múltipla:

H_0: $\beta_1 = \beta_2 = \ldots = \beta_p = 0$
H_a: Um ou mais dos parâmetros não são iguais a zero

Se H_0 for rejeitada, o teste nos dá suficientes evidências estatísticas para concluirmos que um ou mais dos parâmetros não são iguais a zero e que a relação global entre *y* e o conjunto de variáveis independentes x_1, $x_2, \ldots, x_p$ é significativa. Entretanto, se H_0 não puder ser rejeitada, não teremos evidências suficientes para concluir que uma relação significativa está presente.

Antes de descrevermos as etapas do teste *F*, precisamos rever o conceito de *quadrado médio*. Um quadrado médio é a soma dos quadrados dividida por seus graus de liberdade correspondentes. No caso da regressão múltipla, a soma total dos quadrados tem $n - 1$ graus de liberdade, a soma dos quadrados da regressão (SSR) tem *p* graus de liberdade, e a soma dos quadrados dos erros tem $n - p - 1$ graus de liberdade. Portanto, a regressão média quadrática (MSR) é igual a SSR/*p* e o quadrado médio devido aos erros (MSE) é SSE/$(n - p - 1)$.

$$\text{MSR} = \frac{\text{SSR}}{p} \qquad \textbf{(13.12)}$$

e

$$\text{MSE} = \frac{\text{SSE}}{n - p - 1} \qquad \textbf{(13.13)}$$

Conforme discutimos no Capítulo 12, a MSE fornece uma estimativa sem viés de σ^2, que é a variância do termo de erro ϵ. Se H_0:$\beta_1 = \beta_2 = \ldots = \beta_p = 0$ for verdadeira, MSR também fornecerá uma estimativa de σ^2, e o valor de MSR/MSE deve estar próximo de 1. Entretanto, se H_0 for falsa, MSR superestimará σ^2, e o valor de MSR/MSE se tornará maior. Para determinar qual tamanho MSR/MSE deve ter para rejeitarmos H_0, recorremos ao fato de que se H_0 for verdadeira e as suposições sobre o modelo de regressão múltipla forem válidas, a distribuição amostral de MSR/MSE será uma distribuição *F* com *p* graus de liberdade no numerador e $n - p - 1$ no denominador. Um resumo do teste *F* de significância na regressão múltipla é apresentado a seguir.

TESTE *f* DE SIGNIFICÂNCIA GLOBAL

H_0: $\beta_1 = \beta_2 = \ldots = \beta_p = 0$
H_a: Um ou mais dos parâmetros não são iguais a zero

ESTATÍSTICA DE TESTE

$$F = \frac{\text{MSR}}{\text{MSE}} \qquad \textbf{(13.14)}$$

REGRA DE REJEIÇÃO

Critério do valor *p*: Rejeitar H_0 se o valor $p \leq \alpha$
Critério do valor crítico: Rejeitar H_0 se $F \geq F_\alpha$

em que F_α baseia-se em uma distribuição *F* com *p* graus de liberdade no numerador e $n - p - 1$ graus de liberdade no denominador.

Apliquemos o teste *F* ao problema de regressão múltipla da Butler Trucking Company. Com duas variáveis independentes, as hipóteses são escritas da seguinte maneira:

H_0: $\beta_1 = \beta_2 = 0$
H_a: β_1 e/ou β_2 não são iguais a zero

A Figura 13.6 é uma saída de dados do Minitab referente ao modelo de regressão múltipla, tendo as milhas percorridas (x_1) e o número de entregas (x_2) como as variáveis independentes. Na parte da análise de

variância da saída de computador, notamos que MSR = 10,9 e MSE = 0,328. Usando a Equação 13.14, obtemos a estatística de teste.

$$F = \frac{10,8}{0,328} = 32,9$$

Figura 13.6 Saída de dados do Minitab para o problema da Butler Trucking com duas variáveis independentes: as milhas percorridas (x_1) e o número de entregas (x_2)

```
The regression equation is
Time = - 0.869 + 0.0611 Miles + 0.923 Deliveries

Predictor        Coef    SE Coef       T      p
Constant      -0.8687     0.9515   -0.91  0.392
Miles        0.061135   0.009888    6.18  0.000
Deliveries     0.9234     0.2211    4.18  0.004

S = 0.5731    R-sq = 90.4%    R-sq(adj) = 87.6%

Analysis of Variance

SOURCE          DF      SS       MS       F      p
Regression       2  21.601   10.800   32.88  0.000
Residual Error   7   2.299    0.328
Total            9  23.900
```

Observe que o valor F na saída do Minitab é $F = 32,88$; o valor que calculamos difere porque usamos valores arredondados para MSR e MSE no cálculo. Usando $\alpha = 0,01$, o valor $p = 0,0000$ na última coluna da tabela de análise de variância (Figura 13.6) indica que podemos rejeitar H_0: $\beta_1 = \beta_2 = 0$ porque o valor p é menor que $\alpha = 0,01$. Alternativamente, a Tabela 4 do Apêndice B mostra que com dois graus de liberdade no numerador e sete graus de liberdade no denominador, $F_{0,01} = 9,55$. Com $32,9 > 9,55$, rejeitamos H_0: $\beta_1 = \beta_2 = 0$ e concluímos que há uma relação significativa entre o tempo de viagem y e as duas variáveis independentes, as milhas percorridas e o número de entregas.

Conforme observamos anteriormente, o erro médio quadrático fornece uma estimativa sem viés de σ^2, que é a variância do termo de erro ϵ. Consultando a Figura 13.6, notamos que a estimativa de σ^2 é MSE $= 0,328$. A raiz quadrada de MSE é a estimativa do desvio padrão do termo de erro. Conforme definimos na Seção 12.5, esse desvio padrão é chamado de erro padrão da estimativa e é designado por s. Portanto, $s = \sqrt{\text{MSE}} = \sqrt{0,328} = 0,573$. Note que o valor do erro padrão da estimativa aparece na saída do Minitab da Figura 13.6.

A Tabela 13.3 é a tabela de análise de variância (ANOVA) geral que fornece os resultados do teste F de um modelo de regressão múltipla. O valor da estatística de teste F aparece na última coluna e pode ser comparado a F_α com p graus de liberdade no numerador e $n - p - 1$ graus de liberdade no denominador para se tomar a conclusão do teste de hipótese. Ao revisar a saída do Minitab referente à Butler Trucking Company da Figura 13.6, notamos que a tabela de análise de variância do Minitab contém essa informação. Além disso, o Minitab também fornece o valor p correspondente à estatística de teste F.

Tabela 13.3 Tabela ANOVA de um modelo de regressão múltipla com p variáveis independentes

Fonte	Soma dos Quadrados	Graus de Liberdade	Quadrado Médio	F
Regressão	SSR	p	$\text{MSR} = \frac{\text{SSR}}{p}$	$F = \frac{\text{MSR}}{\text{MSE}}$
Erro	SSE	$n - p - 1$	$\text{MSE} = \frac{\text{SSE}}{n - p - 1}$	
Total	SST	$n - 1$		

Teste t

Se o teste F demonstrar que a relação de regressão múltipla é significativa, um teste t pode ser realizado para determinar a significância de cada um dos parâmetros individuais. O teste t de significância individual é o seguinte:

TESTE t DE SIGNIFICÂNCIA INDIVIDUAL
Para qualquer parâmetro β_i

$$H_0: \beta_i = 0$$
$$H_a: \beta_i \neq 0$$

ESTATÍSTICA DE TESTE

$$t = \frac{b_i}{s_{b_i}} \quad \textbf{(13.15)}$$

REGRA DE REJEIÇÃO

Critério do valor p: Rejeitar H_0 se o valor $p \leq \alpha$
Critério do valor crítico: Rejeitar H_0 se $t \leq -t_{\alpha/2}$ ou se $t \geq t_{\alpha/2}$

em que $t_{\alpha/2}$ baseia-se em uma distribuição t com $n - p - 1$ graus de liberdade.

Na estatística de teste, s_{b_i} é a estimativa do desvio padrão de b_i. O valor de s_{b_i} será fornecido pelo software.

Vamos realizar o teste t do problema de regressão da Butler Trucking. Consulte a parte da Figura 13.6 que apresenta a saída de dados do Minitab correspondente aos cálculos da razão t. Os valores de b_1, b_2, s_{b_i} e s_{b_2} são os seguintes:

$$b_1 = 0{,}061135 \quad s_{b_1} = 0{,}009888$$
$$b_2 = 0{,}9234 \quad s_{b_2} = 0{,}2211$$

Usando a Equação 13.15, obtemos a estatística de teste das hipóteses que envolvem os parâmetros β_1 e β_2.

$$t = 0{,}061135/0{,}009888 = 6{,}18$$
$$t = 0{,}9234/0{,}2211 = 4{,}18$$

Note que ambos os valores da razão t e os correspondentes valores p são fornecidos pela saída de dados do Minitab na Figura 13.6. Usando $\alpha = 0{,}01$, os valores p 0,000 e 0,004 apresentados pelo Minitab indicam que podemos rejeitar H_0: $\beta_1 = 0$ e H_0: $\beta_2 = 0$. Portanto, ambos os parâmetros são estatisticamente significativos. Alternativamente, a Tabela 2 do Apêndice B demonstra que com $n - p - 1 = 10 - 2 - 1 = 7$ graus de liberdade, $t_{0,005} = 3{,}499$. Com $6{,}18 > 3{,}499$, rejeitamos H_0: $\beta_1 = 0$. Similarmente, com $4{,}18 > 3{,}499$, rejeitamos H_0: $\beta_2 = 0$.

Multicolinearidade

Utilizamos o termo *variável independente* na análise de regressão ao fazermos referência a qualquer variável que é usada para prever ou explicar o valor da variável dependente. O termo não significa, entretanto, que as variáveis independentes sejam, em si mesmas, independentes no sentido estatístico. Ao contrário, a maioria das variáveis independentes de um problema de regressão múltipla estão, até certo ponto, correlacionadas. Por exemplo, no caso da Butler Trucking envolvendo duas variáveis independentes x_1 (milhas percorridas) e x_2 (número de entregas), poderemos tratar as milhas percorridas como a variável dependente e o número de entregas como a variável independente para determinar se essas duas mesmas variáveis estão relacionadas entre si.

Poderíamos, então, calcular o coeficiente de correlação da amostra, $r_{x_1x_2}$ para determinar o grau em que as variáveis estão relacionadas. Esse cálculo produz $r_{x_1x_2} = 0{,}16$. Desse modo, encontramos certo grau de associação linear entre as duas variáveis independentes. Na análise de regressão múltipla, o termo **multicolinearidade** refere-se à correlação entre as variáveis independentes.

Para oferecermos uma perspectiva melhor dos potenciais problemas da multicolinearidade, consideremos uma modificação no exemplo da Butler Trucking. Em vez de x_2 ser o número de entregas, admitamos que x_2 denote o número de galões de combustível consumidos. Evidentemente, x_1 (as milhas percorridas) e x_2 estão relacionadas; ou seja, sabemos que o número de galões de gasolina utilizados depende do núme-

ro de milhas percorridas. Portanto, poderíamos concluir logicamente que x_1 e x_2 são variáveis independentes altamente correlacionadas.

Suponha obtermos a equação $\hat{y} = b_0 + b_1x_1 + b_2x_2$ e descobrirmos que o teste F demonstra que a relação é significativa. Suponha então realizarmos um teste t em β_1 para determinar se $\beta_1 \neq 0$ e não possamos rejeitar H_0: $\beta_1 = 0$. Esse resultado significa que o tempo de viagem não está relacionado com as milhas percorridas? Não necessariamente. O que provavelmente significa é que, com x_2 já estando no modelo, x_1 não contribui significativamente para determinar o valor de y. Essa interpretação faz sentido em nosso exemplo; se soubermos a quantidade de gasolina consumida, não obtemos muita informação adicional útil para prever y ao sabermos o número de milhas percorridas. Similarmente, um teste t poderia levar-nos a concluir que $\beta_2 = 0$ considerando que, com x_1 no modelo, saber qual é a quantidade de gasolina consumida não nos ajuda muito.

Quando as variáveis independentes são altamente correlacionadas, não é possível determinar o efeito distinto de qualquer variável independente em particular sobre a variável dependente.

Para resumir, em testes t da significância de parâmetros individuais, a dificuldade provocada pela multicolinearidade baseia-se no fato de que é possível concluir que nenhum dos parâmetros individuais é significativamente diferente de zero quando um teste F sobre a equação de regressão múltipla individual indica uma relação significativa. Esse problema é evitado quando há pouca correlação entre as variáveis independentes.

Os estatísticos desenvolveram diversos testes para determinar se a multicolinearidade é suficientemente elevada para causar problemas. De acordo com o teste prático, a multicolinearidade constitui um problema potencial se o valor absoluto do coeficiente de correlação da amostra ultrapassar 0,70 em qualquer das duas variáveis independentes. Os outros tipos de teste são mais avançados e estão além do escopo deste livro.

Um coeficiente de correlação amostral maior que +0,70 ou menor que −0,70 para duas variáveis independentes é um aviso prático de que há potenciais problemas com a multicolinearidade.

Se possível, deve-se tentar evitar incluir variáveis independentes que sejam altamente correlacionadas. Na prática, entretanto, raramente é possível seguir estritamente essa norma. Quando os tomadores de decisão têm motivos para acreditar na presença de uma multicolinearidade substancial, eles precisam perceber que é difícil separar os efeitos das variáveis independentes individuais da variável dependente.

NOTAS E COMENTÁRIOS

Costumeiramente, a multicolinearidade não afeta a maneira pela qual executamos nossa análise de regressão ou interpretamos o resultado de um estudo. Entretanto, quando a multicolinearidade é grave – ou seja, quando duas ou mais das variáveis independentes são altamente correlacionadas –, podemos ter dificuldade para interpretar os resultados dos testes t sobre os parâmetros individuais. Além do tipo de problema ilustrado nesta seção, casos graves de multicolinearidade têm demonstrado que resultam em estimativas pelo método dos mínimos quadrados que tem o sinal errado. Ou seja, em estudos simulados em que os pesquisadores criaram o modelo de regressão subjacente e depois aplicaram a técnica dos mínimos quadrados para desenvolver estimativas de β_0, β_1, β_2 etc., foi demonstrado que, sob condições de elevada multicolinearidade, as estimativas pelo método dos mínimos quadrados podem ter um sinal oposto ao do parâmetro que é estimado. Por exemplo, β_2 poderia ser, de fato, + 10 e b_2, por sua vez, poderia vir a ser −2. Desse modo, não se pode acreditar muito nos coeficientes individuais se houver a presença de multicolinearidade em grau elevado.

Exercícios

Métodos

19. No exemplo 1, foi apresentada a seguinte equação de regressão estimada baseada em dez observações:

$$\hat{y} = 29{,}1270 + 0{,}5906x_1 + 0{,}4980x_2$$

AUTOTESTE

Aqui, SST = 6724,125, SSR = 6216,375, $s_{b_1} = 0{,}0813$ e $s_{b_2} = 0{,}0567$.

a. Calcule MSR e MSE.
b. Calcule F e execute o teste F apropriado. Use $\alpha = 0{,}05$.
c. Realize um teste t da significância de β_1. Use $\alpha = 0{,}05$.
d. Realize um teste t da significância de β_2. Use $\alpha = 0{,}05$.

20. Consulte os dados apresentados no exercício 2. A equação de regressão estimada desses dados é:

$$\hat{y} = 18{,}4 + 2{,}01x_1 + 4{,}47x_2$$

Aqui, SST = 15.182,9, SSR = 14.052,2, s_{b_1} = 0,2471 e s_{b_2} = 0,9484.

a. Teste a relação de significância entre x_1, x_2 e y. Use $\alpha = 0{,}05$.
b. β_1 é significativo? Use $\alpha = 0{,}05$.
c. β_2 é significativo? Use $\alpha = 0{,}05$.

21. A seguinte equação de regressão estimada foi desenvolvida para um modelo que envolve duas variáveis independentes:

$$\hat{y} = 40{,}7 + 8{,}63x_1 + 2{,}71x_2$$

Depois que x_2 foi retirado do modelo, o método dos mínimos quadrados foi usado para obter uma equação de regressão estimada que envolve somente x_1 como a variável independente:

$$\hat{y} = 42{,}0 + 9{,}01x_1$$

a. Apresente uma interpretação do coeficiente de x_1 em ambos os modelos.
b. A multicolinearidade poderia explicar por que o coeficiente de x_1 difere nos dois modelos? Se assim for, como isso acontece?

Aplicações

22. No exercício 4, foi apresentada a seguinte equação de regressão estimada relacionando as vendas com o investimento em estoques e os gastos em propaganda:

$$\hat{y} = 25 + 10x_1 + 8x_2$$

Os dados usados para desenvolver o modelo foram extraídos de uma pesquisa de dez lojas; para esses dados, SST = 16.000 e SSR = 12.000.
a. Calcule SSE, MSE e MSR.
b. Use um teste F e o nível de significância 0,05 para determinar se há uma relação entre as variáveis.

AUTOTESTE

23. Consulte o exercício 5.
a. Use $\alpha = 0{,}01$ pra testar as hipóteses:

$$H_0: \beta_1 = \beta_2 = 0$$
$$H_a: \beta_1 \text{ e/ou } \beta_2 \text{ não são iguais a zero.}$$

para o modelo $y = b_0 + b_1x_1 + b_2x_2 + \epsilon$, em que

$$x_1 = \text{propaganda de televisão (US\$ 1.000)}$$
$$x_2 = \text{propaganda de jornal (US\$ 1.000)}$$

ARQUIVO DA INTERNET Showtime

b. Use $\alpha = 0{,}05$ para testar a significância de β_1. x_1 deve ser retirado do modelo?
c. Use $\alpha = 0{,}05$ para testar a significância de β_2. x_2 deve ser retirado do modelo?

24. Consulte os dados do exercício 6. Use o número de *home runs* feitos pela equipe e a média de *runs* recebidos da equipe arremessadora para prever a porcentagem de jogos ganhos.
a. Use o teste F para determinar a significância geral da relação. Qual é a sua conclusão ao nível de significância 0,05?
b. Use o teste t para determinar a significância de cada variável independente. Qual é a sua conclusão ao nível de significância 0,05?

ARQUIVO DA INTERNET MLB

25. A *Barron's* realiza uma revisão anual das corretoras on-line, incluindo tanto as corretoras que podem ser acessadas por meio de um navegador de internet como as corretoras de acesso direto, as quais conectam os clientes diretamente com o servidor de rede da corretora. As ofertas e o desempenho de cada corretora são avaliados em seis áreas usando uma pontuação de 0 a 5 em cada categoria. Os resultados são ponderados para se obter uma pontuação global e, então, uma classificação final designada por estrelas, a qual varia de zero a cinco estrelas, é atribuída a cada corretora. A execução do negócio, facilidade de uso e a variedade de ofertas são três das áreas avaliadas. Uma pontuação igual a 5 na execução do negócio significa que o processo de entrada e execução do pedido fluiu facilmente de uma etapa para a seguinte. Um valor igual a 5 para a facilidade de uso significa que o site foi fácil de usar e que pode ser personalizado para exibir aquilo que o cliente quer ver. Um valor igual a 5 para a área

de variedade de ofertas significa que todas as transações de investimentos podem ser executadas on-line. Os dados a seguir apresentam as pontuações correspondentes à execução do negócio, facilidade de uso e variedade de ofertas, bem como uma classificação por estrelas de uma amostra de dez das corretoras on-line que a *Barron's* avaliou (*Barron's*, 10 de março de 2003).

Corretora	Execução do Negócio	Facilidade de Uso	Variedade de Ofertas	Avaliação
Wall St. Access	3,7	4,5	4,8	4,0
E*TRADE (Power)	3,4	3,0	4,2	3,5
E*TRADE (Standard)	2,5	4,0	4,0	3,5
Preferred Trade	4,8	3,7	3,4	3,5
my Track	4,0	3,5	3,2	3,5
TD Waterhouse	3,0	3,0	4,6	3,5
Brown & Co.	2,7	2,5	3,3	3,0
Brokerage America	1,7	3,5	3,1	3,0
Merrill Lynch Direct	2,2	2,7	3,0	2,5
Strong Funds	1,4	3,6	2,5	2,0

ARQUIVO DA INTERNET
Brokers

a. Determine a equação de regressão estimada que possa ser usada para prever a classificação por estrelas, dadas as pontuações para a execução, facilidade de uso e variedade de ofertas.
b. Use o teste F para determinar a significância global da relação. Qual é a sua conclusão no nível de significância 0,05?
c. Use o teste t para determinar a significância de cada variável independente. Qual é a sua conclusão ao nível de significância 0,05?
d. Retire da equação de regressão estimada quaisquer variáveis independentes que não sejam significativas. Qual é a sua equação de regressão estimada recomendada? Compare R^2 com o valor de R^2 obtido no item (a). Discuta as diferenças.

26. No exercício 10, foi desenvolvida uma equação de regressão estimada relacionando os jogos ganhos com a porcentagem de *field goals* feitos pela equipe, a porcentagem de lances de três pontos feitos pelo time adversário, e o número de *turnovers* cometidos pelo time adversário.
    a. Use o teste F para determinar a significância global da relação. Qual é a sua conclusão no nível de significância 0,05?
    b. Use o teste t para determinar a significância de cada variável independente. Qual é a sua conclusão no nível de significância 0,05?

ARQUIVO DA INTERNET
NBA

13.6 USANDO A EQUAÇÃO DE REGRESSÃO ESTIMADA PARA ESTIMAÇÃO E PREVISÃO

Os procedimentos para estimar o valor médio de y e para prever um valor individual de y na regressão múltipla são similares aos da análise de regressão que envolvem uma variável independente. Primeiramente, lembre-se de que mostramos no Capítulo 12 que a estimação por ponto do valor esperado de y para determinado valor de x era idêntica à estimação por ponto de um valor individual de y. Em ambos os casos, usamos $\hat{y} = b_0 + b_1x$ como estimação por ponto.

Na regressão múltipla, usamos o mesmo procedimento. Ou seja, substituímos os valores dados, x_1, x_2, . . ., x_p na equação de regressão estimada e usamos o valor correspondente de $\hat{y}$ como estimação por ponto. Suponha que no exemplo da Butler Trucking queiramos usar a equação de regressão estimada envolvendo x_1 (milhas percorridas) e x_2 (número de entregas) para desenvolver duas estimações por intervalo:

1. Um *intervalo de confiança* do tempo médio de viagem para todos os caminhões que percorrem 100 milhas e fazem duas entregas.
2. Um *intervalo de previsão* do tempo de viagem de *um caminhão específico* que percorre 100 milhas e faz duas entregas.

Usando a equação de regressão estimada $\hat{y} = -0{,}869 + 0{,}0611x_1 + 0{,}923x_2$, sendo $x_1 = 100$ e $x_2 = 2$, obtemos o seguinte valor de $\hat{y}$:

$$\hat{y} = -0,869 + 0,0611(100) + 0,923(2) = 7,09$$

Portanto, a estimação por ponto do tempo de viagem em ambos os casos é de aproximadamente sete horas.

Para desenvolver estimações por intervalo do valor médio de y e de um valor individual de y, usamos um procedimento similar ao da análise de regressão que envolve uma variável independente. As fórmulas necessárias estão além do escopo deste livro, mas softwares freqüentemente fornecem intervalos de confiança tão logo os valores de $x_1, x_2, \ldots, x_p$ são especificados pelo usuário. Na Tabela 13.4, apresentamos os intervalos de confiança e de previsão de 95% do exemplo da Butler Company correspondentes a valores selecionados de x_1 e x_2; esses valores foram obtidos com o Minitab. Note que a estimação por intervalo de um valor individual de y é mais ampla que a estimação por intervalo do valor esperado de y. Essa diferença simplesmente reflete o fato de podermos estimar o tempo médio de viagem de todos os caminhões, considerando determinados valores de x_1 e x_2, do que podemos prever o tempo de viagem de um caminhão específico.

Tabela 13.4 Os intervalos de confiança e de previsão de 95% da Butler Trucking

Valor de x_1	Valor de x_2	Intervalo de Confiança		Intervalo de Previsão	
		Limite Mínimo	Limite Máximo	Limite Mínimo	Limite Máximo
50	2	3,146	4,924	2,414	5,656
50	3	4,127	5,789	3,368	6,548
50	4	4,815	6,948	4,157	7,607
100	2	6,258	7,926	5,500	8,683
100	3	7,385	8,645	6,520	9,510
100	4	8,135	9,742	7,362	10,515

Exercícios

Métodos

27. No exercício 1, foi apresentada a seguinte equação de regressão estimada baseada em dez observações:

$$\hat{y} = 29,1270 + 0,5960x_1 + 0,4980x_2$$

a. Desenvolva uma estimação por ponto do valor médio de y quando $x_1 = 180$ e $x_2 = 310$.
b. Desenvolva uma estimação por ponto de um valor individual de y quando $x_1 = 180$ e $x_2 = 310$.

AUTOTESTE

28. Consulte os dados do exercício 2. A equação de regressão estimada desses dados é:

$$\hat{y} = -18,4 + 2,01x_1 + 4,74x_2$$

a. Desenvolva um intervalo de confiança de 95% para o valor médio de y quando $x_1 = 45$ e $x_2 = 15$.
b. Desenvolva um intervalo de previsão de 95% para y quando $x_1 = 45$ e $x_2 = 15$.

Aplicações

AUTOTESTE

ARQUIVO DA INTERNET Showtime

29. No exercício 5, o proprietário da Showtime Movie Theaters, Inc., usou análise de regressão múltipla para prever a receita bruta (y) em função da propaganda de televisão (x_1) e da propaganda de jornal (x_2). A equação de regressão estimada foi:

$$\hat{y} = 83,2 + 2,29x_1 + 1,30x_2$$

a. Qual é a receita bruta esperada de uma semana quando US$ 3.500 foram gastos em propaganda de televisão ($x_1 = 3,5$) e US$ 1.800 foram gastos em propaganda de jornal ($x_2 = 1,8$)?
b. Forneça um intervalo de confiança de 95% correspondente à receita média de todas as semanas que apresentaram os gastos relacionados no item (a).
c. Forneça um intervalo de confiança de 95% correspondente à receita da próxima semana, supondo que os gastos de propaganda serão alocados como no item (a).

30. No exercício 9, foi desenvolvida uma equação de regressão estimada relacionando a porcentagem de estudantes que freqüentam um curso superior de quatro anos com o tamanho médio da classe e a pontuação média no SAT.

ARQUIVO DA INTERNET **Schools**

   a. Desenvolva um intervalo de confiança de 95% correspondente à porcentagem média dos estudantes que freqüentam um curso superior de quatro anos em um distrito escolar que tem um tamanho médio de classe igual a 25 e cujos estudantes têm uma pontuação média no SAT igual a 1.000.
   b. Suponha que um distrito escolar da cidade de Conway, na Carolina do Sul, tenha um tamanho médio de classe igual a 25 e uma pontuação média no SAT igual a 950. Desenvolva um intervalo de previsão de 95% da porcentagem de estudantes que freqüentam um curso superior de quatro anos.

31. A seção Buyer's Guide (Guia do Comprador) do site da revista *Car and Drive* fornece avaliações e testes de estrada de carros, caminhões, utilitários esportivos e vans. A média das avaliações da qualidade geral, estilo do veículo, freios, manejo, economia de combustível, conforto interno, aceleração, confiabilidade, ajuste e acabamento, transmissão e tração de cada veículo é resumida usando-se uma escala que varia de 1 (o pior) a 10 (o melhor). Uma parte dos dados referentes a 14 carros esportivos/GT é apresentada a seguir (http://www.caranddriver, 7 de janeiro de 2004).

ARQUIVO DA INTERNET **SportsCar**

Esportivo/GT	Avaliação Geral	Manejo	Confiabilidade	Ajuste e Acabamento
Acura 3.2CL	7,80	7,83	8,17	7,67
Acura RSX	9,02	9,46	9,35	8,97
Audi TT	9,00	9,58	8,74	9,38
BMW 3-Series/M3	8,39	9,52	8,39	8,55
Chevrolet Corvette	8,82	9,64	8,54	7,87
Ford Mustang	8,34	8,85	8,70	7,34
Honda Civic Si	8,92	9,31	9,50	7,93
Infinity G35	8,70	9,34	8,96	8,07
Mazda RX-8	8,58	9,79	8,96	8,12
Mini Cooper	8,76	10,00	8,69	8,33
Mitsubishi Eclipse	8,17	8,95	8,25	7,36
Nissan 350Z	8,07	9,35	7,56	8,21
Porsche 911	9,55	9,91	8,86	9,55
Toyota Celica	8,77	9,29	9,04	7,97

   a. Desenvolva uma equação de regressão estimada usando a capacidade de manejo, confiabilidade e ajuste e acabamento para prever a qualidade geral.
   b. Outro carro esportivo/GT avaliado pela *Car and Drive* é o Honda Accord. As avaliações de manejo, confiabilidade e ajuste e acabamento do Honda Accord foram 8,28, 9,06 e 8,07, respectivamente. Estime a classificação geral desse carro.
   c. Forneça um intervalo de confiança de 95% da qualidade geral de todos os carros esportivos e GT com as características relacionadas no item (b).
   d. Forneça um intervalo de previsão de 95% da qualidade geral do Honda Accord descrito no item (b).
   e. A avaliação geral divulgada pela *Car and Drive* para o Honda Accord foi 8,65. Como essa avaliação se compara com as estimativas que você desenvolveu nos itens (b) e (d)?

13.7 VARIÁVEIS QUALITATIVAS INDEPENDENTES

As variáveis independentes podem ser qualitativas ou quantitativas.

Até aqui, os exemplos que consideramos envolveram variáveis quantitativas independentes, como a população de estudantes, a distância percorrida e o número de entregas. Em muitas situações, entretanto, devemos trabalhar com **variáveis qualitativas independentes**, como o sexo (masculino, feminino), método de pagamento (dinheiro, cartão de crédito, cheque) e assim por diante. O propósito desta seção é mostrar-lhe como as variáveis qualitativas são tratadas na análise de regressão. Para ilustrar o uso e a interpretação de uma variável qualitativa independente, consideraremos um problema enfrentado pelos gerentes da Johnson Filtration, Inc.

Exemplo: Johnson Filtration, Inc.

A Johnson Filtration, Inc. oferece serviços de manutenção para sistemas de filtração de água em todo o sul da Flórida. Os clientes contatam a Johnson solicitando serviços de manutenção em seus sistemas de filtra-

gem de água. Para estimar o tempo de atendimento e o custo do serviço, os gerentes da Johnson querem prever o tempo de reparo necessário para cada pedido de manutenção. Portanto, o tempo de reparo em horas é a variável dependente. Acredita-se que o tempo de reparo esteja relacionado a dois fatores: o número de meses desde o último serviço de manutenção e o tipo de problema que requer o reparo (mecânico ou elétrico). Os dados de uma amostra de dez chamadas de serviço estão registrados na Tabela 13.5.

Digamos que y denote o tempo de reparo em horas e que x_1 designe o número de meses desde o último serviço de manutenção. O modelo de regressão que usa somente x_1 para prever y é:

$$y = \beta_0 + \beta_1 x_1 + \epsilon$$

Usando o Minitab para desenvolver a equação de regressão estimada, obtivemos a saída mostrada na Figura 13.7. A equação de regressão estimada é:

$$\hat{y} = 2{,}15 + 0{,}304x_1 \qquad \textbf{(13.16)}$$

Tabela 13.5 Dados do exemplo da Johnson Filtration

Chamada de Serviço	Número de Meses Desde o Último Serviço	Tipo de Conserto	Tempo de Reparo em Horas
1	2	Elétrico	2,9
2	6	Mecânico	3,0
3	8	Elétrico	4,8
4	3	Mecânico	1,8
5	2	Elétrico	2,9
6	7	Elétrico	4,9
7	9	Mecânico	4,2
8	8	Mecânico	4,8
9	4	Elétrico	4,4
10	6	Elétrico	4,5

No nível de significância de 0,05, o valor p igual a 0,016 para o teste t (ou F) indica que o número de meses desde o último serviço de manutenção está significativamente relacionado com o tempo de reparo. R-sq = 53,4% indica que x_1 isoladamente explica 53,4% da variabilidade no tempo de reparo.

Para incorporar o tipo de reparo no modelo de regressão, definimos a seguinte variável:

$$x_2 \begin{cases} 0 \text{ se o tipo de reparo for mecânico} \\ 1 \text{ se o tipo de reparo for elétrico} \end{cases}$$

Na análise de regressão, x_2 é chamada **variável** (ou *indicador*) **simulada**. Usando essa variável simulada, podemos escrever o modelo de regressão múltipla como:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon$$

A Tabela 13.6 é o conjunto de dados (*data set*) revisado que inclui os valores da variável simulada. Usando o Minitab e os dados da Tabela 13.6, podemos desenvolver estimativas dos parâmetros do modelo. A saída do Minitab apresentada na Figura 13.8 mostra que a equação de regressão múltipla estimada é

$$\hat{y} = 0{,}93 + 0{,}388x_1 + 1{,}26x_2 \qquad \textbf{(13.17)}$$

Figura 13.7 Saída do Minitab referente à Johnson Filtration, tendo como variável independente o número de meses desde o último serviço de manutenção (x_1)

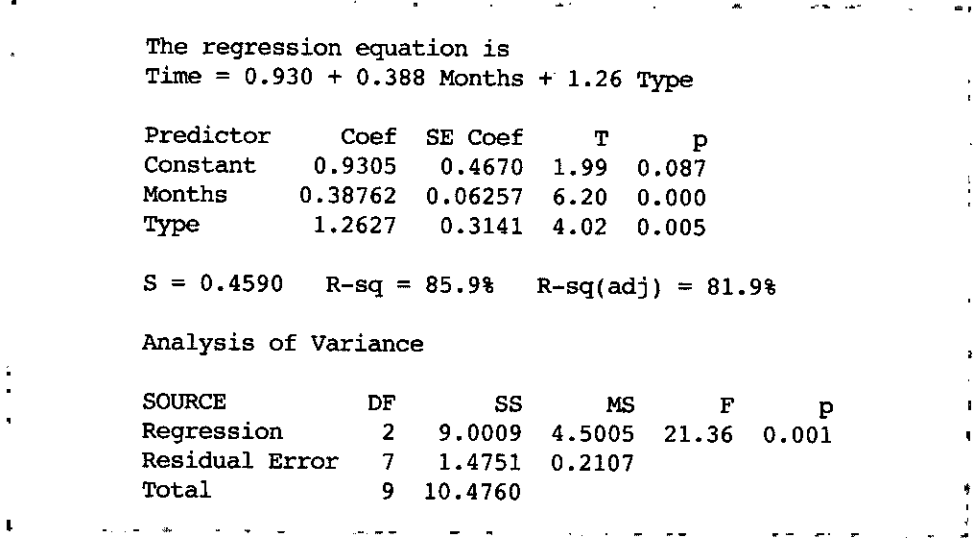
```
The regression equation is
Time = 0.930 + 0.388 Months + 1.26 Type

Predictor      Coef   SE Coef     T      p
Constant     0.9305    0.4670  1.99  0.087
Months      0.38762   0.06257  6.20  0.000
Type         1.2627    0.3141  4.02  0.005

S = 0.4590    R-sq = 85.9%    R-sq(adj) = 81.9%

Analysis of Variance

SOURCE          DF       SS      MS      F      p
Regression       2   9.0009  4.5005  21.36  0.001
Residual Error   7   1.4751  0.2107
Total            9  10.4760
```

Na saída do Minitab, os nomes das variáveis *Months* e *Time* foram inseridos na planilha como cabeçalhos de coluna; desse modo, x_1 = *Months* e y = *Time*.

Tabela 13.6 Dados do exemplo da Johnson Filtration, sendo o tipo de reparo indicado por uma variável simulada ($x_2 = 0$ para reparos mecânicos e $x_2 = 1$ para reparos elétricos)

Cliente	Número de Meses Desde o Último Serviço (x_1)	Tipo de Conserto (x_2)	Tempo de Reparo em Horas (y)
1	2	1	2,9
2	6	0	3,0
3	8	1	4,8
4	3	0	1,8
5	2	1	2,9
6	7	1	4,9
7	9	0	4,2
8	8	0	4,8
9	4	1	4,4
10	6	1	4,5

ARQUIVO DA INTERNET
Johnson

No nível de significância 0,05, o valor p igual a 0,001 associado ao teste F ($F = 21{,}36$) indica que a relação de regressão é significativa. A parte do teste t da saída computadorizada da Figura 13.8 indica que tanto os meses desde o último serviço (valor $p = 0{,}000$) como o tipo de reparo (valor $p = 0{,}005$) são estatisticamente significativos. Além disso, R-sq = 85,9% e R-sq(adj) = 81,9% indicam que a equação de regressão estimada explica bem a variabilidade nos tempos de reparo. Desse modo, a Equação 13.17 se demonstrará útil em termos de estimar o tempo de reparo necessário para as várias chamadas de serviço de manutenção.

Interpretando os Parâmetros

A equação de regressão múltipla para o exemplo da Johnson Filtration é:

$$E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 \qquad (13.18)$$

Na saída do Minitab, os nomes das variáveis *Months*, *Type* e *Time* foram inseridos na planilha como cabeçalhos de coluna; desse modo, x_1 = *Months*, x_2 = *Type* e y = *Time*.

Figura 13.8 Saída do Minitab referente à Johnson Filtration, tendo como variáveis independentes o número de meses desde o último serviço de manutenção (x_1) e o tipo de reparo (x_2)

```
The regression equation is
Time = 0.930 + 0.388 Months + 1.26 Type

Predictor      Coef   SE Coef      T      p
Constant     0.9305    0.4670   1.99  0.087
Months      0.38762   0.06257   6.20  0.000
Type         1.2627    0.3141   4.02  0.005

S = 0.4590    R-sq = 85.9%    R-sq(adj) = 81.9%

Analysis of Variance

SOURCE           DF        SS       MS      F      p
Regression        2    9.0009   4.5005  21.36  0.001
Residual Error    7    1.4751   0.2107
Total             9   10.4760
```

Para entender como interpretar os parâmetros β_0, β_1 e β_2 quando uma variável qualitativa está presente, considere o caso em que $x_2 = 0$ (reparo mecânico). Usando $E(y \mid \text{mecânico})$ para designar a média, ou valor esperado, do tempo de reparo *dado* um reparo mecânico, temos:

$$E(y \mid \text{mecânico}) = \beta_0 + \beta_1 x_1 + \beta_2(0) = \beta_0 + \beta_1 x_1 \qquad \textbf{(13.19)}$$

Similarmente, em relação a um reparo elétrico ($x_2 = 1$), temos:

$$\begin{aligned} E(y \mid \text{elétrico}) &= \beta_0 + \beta_1 x_1 + \beta_2(1) = \beta_0 + \beta_1 x_1 + \beta_2 \\ &= (\beta_0 + \beta_2) + \beta_1 x_1 \end{aligned} \qquad \textbf{(13.20)}$$

Ao comparar as Equações 13.19 e 13.20, notamos que o tempo médio de reparo é uma função linear de x_1 tanto para os reparos mecânicos como para os elétricos. A inclinação de ambas as equações é β_1, mas o ponto de interseção com o eixo y difere. Em relação aos reparos mecânicos, o ponto de interseção com y é β_0 na Equação 13.19 e, em relação aos reparos elétricos é, $(\beta_0 + \beta_2)$ na Equação 13.20. A interpretação de β_2 é que ele indica a diferença entre o tempo médio de reparo de problemas elétricos e o tempo médio de reparo para problemas mecânicos.

Se β_2 for positivo, o tempo médio de reparo de um problema elétrico será maior que o de um problema mecânico; se β_2 for negativo, o tempo médio de reparo de um problema elétrico será menor que o de um problema mecânico. Finalmente, se $\beta_2 = 0$, não haverá diferença no tempo médio de reparo de problemas elétricos e mecânicos, e o tipo de reparo não está relacionado com o tempo de reparo.

Usando a equação de regressão múltipla estimada $\hat{y} = 0{,}93 + 0{,}388x_1 + 1{,}26x_2$, notamos que 0,93 é a estimativa de β_0 e 1,26 é a estimativa de β_2. Assim, quando $x_2 = 0$, (reparo mecânico),

$$\hat{y} = 0{,}93 + 0{,}388x_1 \qquad \textbf{(13.21)}$$

e quando $x_2 = 1$ (reparo elétrico),

$$\begin{aligned} \hat{y} &= 0{,}93 + 0{,}388x_1 + 1{,}26(1) \\ &= 2{,}19 + 0{,}388x_1 \end{aligned} \qquad \textbf{(13.22)}$$

De fato, o uso de uma variável simulada para o tipo de reparo produz duas equações que podem ser usadas para prever o tempo de reparo, e uma delas corresponde a reparos mecânicos e a outra se refere a reparos elétricos. Além disso, com $b_2 = 1{,}26$, sabemos que, em média, os reparos elétricos requerem 1,26 horas a mais que os reparos mecânicos.

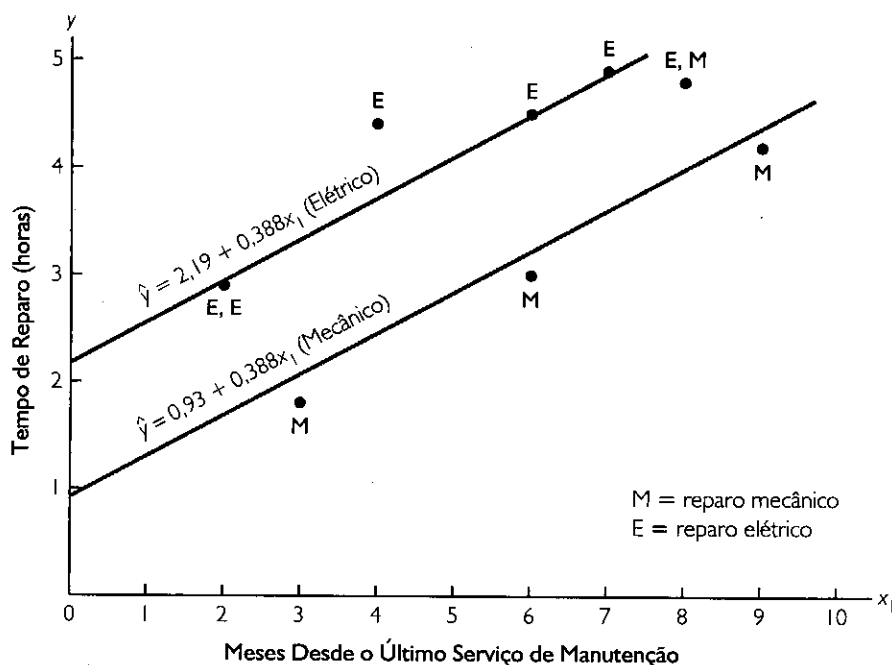
A Figura 13.9 é a plotagem dos dados da Johnson da Tabela 13.6. O tempo de reparo em horas (y) é representado pelo eixo vertical, e os meses desde o último serviço de manutenção (x_1) são representados pelo eixo horizontal. Um ponto de dados correspondente a um reparo mecânico é indicado por um M, e um ponto de dados que se refere a um reparo elétrico é indicado por um E. As Equações 13.21 e 13.22 estão plotadas para exibir graficamente as duas equações que podem ser usadas para prever o tempo de reparo, sendo uma correspondente aos reparos mecânicos e uma, aos reparos elétricos.

Variáveis Qualitativas mais Complexas

Uma variável qualitativa com k níveis deve ser modelada usando-se $k-1$ variáveis simuladas. Deve-se ter cautela tanto ao definir como ao interpretar as variáveis simuladas.

Uma vez que a variável qualitativa do exemplo da Johnson Filtration tinha dois níveis (mecânico e elétrico), foi fácil definirmos uma variável simulada, com 0 (zero) indicando um reparo mecânico e 1, um reparo elétrico. Entretanto, quando uma variável qualitativa tem mais de dois níveis, deve-se ter cautela tanto ao definir como ao interpretar as variáveis simuladas. Conforme mostraremos, se uma variável qualitativa tiver k níveis, $k - 1$ variáveis simuladas serão necessárias, sendo cada variável simulada codificada como 0 ou 1.

Por exemplo, suponha que um fabricante de máquinas copiadoras tenha organizado os territórios de venda de um estado em particular em três regiões: A, B e C. Os gerentes querem usar análise de regressão para ajudar a prever o número de copiadoras vendidas por semana. Sendo o número de unidades vendidas a variável dependente, eles consideram diversas variáveis independentes (o número dos integrantes da equipe de vendas, os gastos de propaganda e assim por diante). Suponha que os gerentes acreditem que a região de vendas também seja um fator importante para preverem o número de copiadoras vendidas. Desde que a região de vendas seja uma variável qualitativa com três níveis, A, B e C, precisaremos de $3 - 1 = 2$ variáveis simuladas para representar a região de vendas.

Figura 13.9 Diagrama de dispersão dos dados de reparos da Johnson Filtration relativos à Tabela 13.6

Cada variável pode ser codificada como 0 ou 1 da seguinte maneira:

$$x_1 = \begin{cases} 1 \text{ se a região de vendas for B} \\ 0 \text{ caso contrário} \end{cases}$$

$$x_2 = \begin{cases} 1 \text{ se a região de vendas for C} \\ 0 \text{ caso contrário} \end{cases}$$

Com essa definição, temos os seguintes valores de x_1 e x_2:

Região	x_1	x_2
A	0	0
B	1	0
C	0	1

As observações correspondentes à região A seriam codificadas como $x_1 = 0$, $x_2 = 0$; as observações correspondentes à região B seriam codificadas como $x_1 = 1$, $x_2 = 0$; e as observações correspondentes à região C seriam codificadas como $x_1 = 0$, $x_2 = 1$.

A equação de regressão relacionando o valor esperado do número de unidades vendidas, $E(y)$, com a variável simulada seria escrita como:

$$E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

Para nos ajudar a interpretar os parâmetros b_0, b_1 e b_2, considere as três variações seguintes da equação de regressão:

$$E(y \mid \text{região A}) = \beta_0 + \beta_1(0) + \beta_2(0) = \beta_0$$
$$E(y \mid \text{região B}) = \beta_0 + \beta_1(1) + \beta_2(0) = \beta_0 + \beta_1$$
$$E(y \mid \text{região C}) = \beta_0 + \beta_1(0) + \beta_2(1) = \beta_0 + \beta_2$$

Desse modo, β_0 é a média, ou valor esperado, das vendas para a região A; β_1 é a diferença entre o número médio de unidades vendidas na região B e o número médio de unidades vendidas na região A; e β_2 é a diferença entre o número médio de unidades vendidas na região C e o número médio de unidades vendidas na região A.

Foram necessárias duas variáveis simuladas porque a região de vendas é uma variável qualitativa com três níveis. Mas a designação de $x_1 = 0$, $x_2 = 0$ para indicar a região A, $x_1 = 1$, $x_2 = 0$ para indicar a região B e $x_1 = 0$, $x_2 = 1$ para indicar a região C foi arbitrária. Por exemplo, poderíamos ter optado por $x_1 = 1$, $x_2 = 0$ para indicar a região A, $x_1 = 0$, $x_2 = 0$ para indicar a região B e $x_1 = 0$, $x_2 = 1$ para indicar a região C. Nesse caso, β_1 teria sido interpretado como a diferença média entre as regiões A e B e β_2 como a diferença média entre as regiões C e B.

O ponto importante a ser lembrado é que, quando uma variável qualitativa tem k níveis, $k - 1$ variáveis simuladas são necessárias na análise de regressão múltipla. Dessa forma, se o exemplo das regiões de venda tivesse uma quarta região, intitulada D, três variáveis simuladas seriam necessárias. Por exemplo, as três variáveis simuladas podem ser codificadas da seguinte maneira:

$$x_1 = \begin{cases} 1 \text{ se a região de vendas for B} \\ 0 \text{ caso contrário} \end{cases} \qquad x_2 = \begin{cases} 1 \text{ se a região de vendas for C} \\ 0 \text{ caso contrário} \end{cases}$$

$$x_3 = \begin{cases} 1 \text{ se a região de vendas for D} \\ 0 \text{ caso contrário} \end{cases}$$

Exercícios

Métodos

AUTOTESTE

32. Considere um estudo de regressão envolvendo uma variável dependente y, uma variável quantitativa independente x_1 e uma variável qualitativa com dois níveis (nível 1 e nível 2).
    a. Escreva uma equação de regressão múltipla relacionando x_1 e a variável qualitativa com y.
    b. Qual é o valor esperado do y correspondente ao nível 1 da variável qualitativa?
    c. Qual é o valor esperado do y correspondente ao nível 2 da variável qualitativa?
    d. Interprete os parâmetros de sua equação de regressão.
33. Considere um estudo de regressão envolvendo uma variável dependente y, uma variável quantitativa independente x_1 e uma variável qualitativa com três níveis possíveis (nível 1, nível 2 e nível 3).
    a. Quantas variáveis simuladas são necessárias para representar a variável qualitativa?
    b. Escreva uma equação de regressão múltipla relacionando x_1 e a variável qualitativa a y.
    c. Interprete os parâmetros de sua equação de regressão.

Aplicações

34. A administração propôs o seguinte modelo de regressão para prever as vendas em uma loja de *fast-food.*

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \epsilon$$

em que

x_1 = o número de concorrentes dentro de uma milha

x_2 = a população dentro de uma milha (em milhares de pessoas)

$$x_3 = \begin{cases} 1 \text{ se houver guichê de vendas } \textit{drive-thru} \\ 0 \text{ caso contrário} \end{cases}$$

y = vendas (em milhares de dólares)

A seguinte equação de regressão estimada foi desenvolvida depois que 20 lojas foram pesquisadas:

$$\hat{y} = 10{,}1 + 4{,}2x_1 + 6{,}8x_2 + 15{,}3x_3$$

a. Qual é o valor de vendas esperado que se pode atribuir ao guichê de vendas *drive-thru*?
b. Preveja as vendas de uma loja com dois concorrentes, uma população de 8 mil habitantes dentro de uma milha e nenhum guichê de vendas *drive-thru.*
c. Preveja as vendas de uma loja com um concorrente, uma população de 3 mil habitantes dentro de uma milha e um guichê de vendas *drive-thru.*

35. Consulte o problema da Johnson Filtration apresentado nesta seção. Suponha que, além da informação sobre o número de meses desde que a máquina sofreu manutenção e se a falha ocorrida foi mecânica ou elétrica, os gerentes obtiveram uma lista indicando qual técnico executou o serviço. Os dados revisados são os seguintes:

Tempo de Reparo em Horas	Meses Desde o Último Serviço de Manutenção	Tipo de Reparo	Técnico
2,9	2	Elétrico	Dave Newton
3,0	6	Mecânico	Dave Newton
4,8	8	Elétrico	Bob Jones
1,8	3	Mecânico	Dave Newton
2,9	2	Elétrico	Dave Newton
4,9	7	Elétrico	Bob Jones
4,2	9	Mecânico	Bob Jones
4,8	8	Mecânico	Bob Jones
4,4	4	Elétrico	Bob Jones
4,5	6	Elétrico	Dave Newton

a. Ignore, por ora, o número de meses desde que ocorreu o último serviço de manutenção (x_1) e o técnico que executou o serviço. Desenvolva uma equação de regressão linear simples estimada para prever o tempo de reparo (y), dado o tipo de reparo (x_2). Lembre-se de que $x_2 = 0$ se o tipo de reparo for mecânico e igual a 1 se o tipo de reparo for elétrico.
b. A equação que você desenvolveu no item (a) proporciona um bom ajuste para os dados observados? Explique.
c. Ignore, por ora, o número de meses desde que ocorreu o último serviço de manutenção e o tipo de reparo associado à máquina. Desenvolva a equação de regressão linear simples estimada para prever o tempo de reparo, dado o técnico que executou o serviço. Admitamos que $x_3 = 0$ se Bob Jones tiver executado o serviço e $x_3 = 1$ se Dave Newton tiver executado o serviço.
d. A equação que você desenvolveu no item (c) proporciona um bom ajuste para os dados observados? Explique.

36. Esse problema é uma extensão da situação descrita no exercício 35.

a. Desenvolva a equação de regressão estimada para prever o tempo de reparo, dado o número de meses desde que ocorreu o último serviço de manutenção, o tipo de reparo e o técnico que executou o serviço.
b. No nível de significância de 0,05, teste se a equação de regressão estimada desenvolvida no item (a) representa uma relação significativa entre as variáveis independentes e a variável dependente.

c. A adição da variável independente x_3, o técnico que executou o serviço, é estatisticamente significativa? Use $\alpha = 0{,}05$. Qual explicação você pode apresentar para os resultados observados?

37. A National Football League avalia os candidatos a jogador de acordo com a posição, em uma escala que varia de 5 a 9. As avaliações são interpretadas da seguinte maneira: 8 a 9 devem começar no primeiro ano; 7,0 a 7,9 estão aptos a começar; 6,0 a 6,9 comporão a equipe como reservas; e 5,0 a 5,9 podem fazer parte do clube e contribuir. A tabela a seguir apresenta a posição, peso, velocidade (para 36,57 m) e as classificações de 25 candidatos à NFL (*USA Today*, 14 de abril, 2000).

ARQUIVO DA INTERNET
Football

Nome	Posição	Peso (kg)	Velocidade (segundos)	Classificação
Cosey Coleman	*Guard*	146,05	5,38	7,4
Travis Claridge	*Guard*	137,43	5,18	7,0
Kaulana Noa	*Guard*	143,78	5,34	6,8
Leander Jordan	*Guard*	149,68	5,46	6,7
Chad Clifton	*Guard*	151,49	5,181	6,3
Manula Savea	*Guard*	139,70	5,32	6,1
Ryan Johanningmeir	*Guard*	140,61	5,28	6,0
Mark Tauscher	*Guard*	144,24	5,37	6,0
Blaine Saipaia	*Guard*	145,60	5,25	6,0
Richard Mercier	*Guard*	133,80	5,34	5,8
Damion McIntosh	*Guard*	148,77	5,31	5,3
Jeno James	*Guard*	145,14	5,64	5,0
Al Jackson	*Guard*	137,89	5,20	5,0
Chris Samuels	*Offensive tackle*	147,41	4,95	8,5
Stockar McDouglas	*Offensive tackle*	163,74	5,50	8,0
Chris McIngosh	*Offensive tackle*	142,88	5,39	7,8
Adrian Klemm	*Offensive tackle*	139,25	4,98	7,6
Todd Wade	*Offensive tackle*	147,87	5,20	7,3
Marvel Smith	*Offensive tackle*	145,15	5,36	7,1
Michael Thompson	*Offensive tackle*	130,18	5,05	6,8
Bobby Williams	*Offensive tackle*	150,59	5,26	6,8
Darnell Alford	*Offensive tackle*	151,50	5,55	6,4
Terrance Beadles	*Offensive tackle*	141,52	5,15	6,3
Tutan Reyes	*Offensive tackle*	135,64	5,35	6,1
Greg Robinson-Ran	*Offensive tackle*	151,05	5,59	6,0

a. Desenvolva uma variável simulada que leve em conta a posição do jogador.
b. Elabore uma equação de regressão estimada para mostrar como a classificação está relacionada com a posição, peso e velocidade.
c. No nível de significância de 0,05, teste se a equação de regressão estimada desenvolvida no item (b) indica uma relação significativa entre as variáveis independentes e a variável dependente.
d. A equação de regressão estimada proporciona um bom ajuste para os dados observados? Explique.
e. A posição é um fator significativo na classificação do jogador? Use $\alpha = 0{,}05$. Explique.
f. Suponha que um novo candidato à posição de *offensive tackle* que pesa 136 kg corra os 36,57 metros (40 jardas) em 5,1 segundos. Use a equação de regressão estimada desenvolvida no item (b) para estimar a classificação desse jogador.

38. Um estudo de anos levado a efeito pela American Heart Association forneceu dados sobre a maneira pela qual a idade, pressão arterial e o tabagismo se relacionam com o risco de acidentes vasculares cerebrais. Suponha que os dados a seguir sejam de uma parte desse estudo. O risco é interpretado como a probabilidade (vezes 100) de o paciente sofrer um derrame cerebral nos próximos dez anos. Em relação à variável tabagismo, defina uma variável simulada com 1 indicando fumante e 0, não-fumante.

ARQUIVO DA INTERNET
Stroke

Risco	Idade	Pressão Arterial	Fumante
12	57	152	Não
24	67	163	Não
13	58	155	Não
56	86	177	Sim
28	59	196	Não
51	76	189	Sim
18	56	155	Sim
31	78	120	Não

Risco	Idade	Pressão Arterial	Fumante
37	80	135	Sim
15	78	98	Não
22	71	152	Não
36	70	173	Sim
15	67	135	Sim
48	77	209	Sim
15	60	199	Não
36	82	119	Sim
8	66	166	Não
34	80	125	Sim
3	62	117	Não
37	59	207	Sim

a. Desenvolva uma equação de regressão estimada que relacione o risco de derrame cerebral com a idade e pressão arterial da pessoa, e se ela é fumante.

b. O tabagismo é um fator significativo no risco de um derrame cerebral? Explique. Use $\alpha = 0{,}05$.

c. Qual é a probabilidade de Art Speen sofrer um derrame cerebral nos próximos dez anos, sendo ele um senhor de 68 anos, fumante, cuja pressão arterial é 175 mmHg?[5] Quais medidas o médico poderia recomendar para esse paciente?

Resumo

Neste capítulo, introduzimos a análise de regressão múltipla como uma extensão da análise de regressão linear simples apresentada no Capítulo 12. A análise de regressão múltipla nos possibilita entender como uma variável dependente se relaciona com duas ou mais variáveis independentes. A equação de regressão múltipla $E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots \beta_p x_p$ indica que o valor esperado, ou valor médio, da variável dependente y está relacionado com os valores das variáveis independentes $x_1, x_2, \ldots, x_p$. Dados amostrais e o método dos mínimos quadrados são usados para desenvolver a equação de regressão múltipla estimada $\hat{y} = b_0 + b_1 x_1 + b_2 x_2 + \ldots + b_p x_p$. Com efeito, $b_0, b_1, b_2, \ldots, b_p$ são estatísticas amostrais usadas para estimar os parâmetros desconhecidos do modelo, $\beta_0, \beta_1, \beta_2, \ldots, \beta_p$. Saídas computadorizadas foram utilizadas ao longo de todo o capítulo para enfatizar o fato de que os softwares estatísticos são o único meio realístico de realizar os numerosos cálculos necessários na análise de regressão múltipla.

O coeficiente de determinação múltiplo foi apresentado como uma medida da eficiência de ajuste da equação de regressão estimada. Ele determina a proporção da variação de y que pode ser explicada pela equação de regressão estimada. O coeficiente de determinação múltiplo ajustado é uma medida similar da eficiência de ajuste que adequa o número de variáveis independentes e, dessa forma, evita superestimar o impacto de acrescentar mais variáveis independentes.

Um teste F e um teste t foram apresentados como maneiras de determinar estatisticamente se a relação entre as variáveis é significativa. O teste F é usado para estabelecer se há uma relação significativa global entre a variável dependente e o conjunto de todas as variáveis independentes. O teste t é usado para determinar se há uma relação significativa entre a variável dependente e uma variável independente individual, dadas as outras variáveis independentes do modelo de regressão. A correlação entre as variáveis independentes, conhecidas como multicolinearidade, também foi discutida.

O capítulo encerrou-se com uma seção sobre como se pode usar variáveis simuladas para incorporar variáveis qualitativas independentes na análise de regressão múltipla.

Glossário

Análise de regressão múltipla Análise de regressão que envolve duas ou mais variáveis independentes.

Modelo de regressão múltipla A equação matemática que descreve como a variável dependente y se relaciona com as variáveis independentes $x_1, x_2, \ldots, x_p$ e um termo de erro ϵ.

[5] NT: mmHG – Milímetros de mercúrio (medida indicada nos aparelhos de medir a pressão arterial).

Equação de regressão múltipla A equação matemática que descreve como a média, ou valor esperado, da variável dependente y se relaciona com os valores das variáveis independentes; ou seja, $E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p$.

Equação de regressão múltipla estimada A estimativa da equação de regressão múltipla baseada em dados amostrais e no método dos mínimos quadrados: $\hat{y} = b_0 + b_1 x_1 + b_2 x_2 + \ldots + b_p x_p$.

Método dos mínimos quadrados O método usado para desenvolver a equação de regressão estimada. Ele minimiza o somatório dos resíduos quadráticos (os desvios entre os valores observados da variável dependente, y_i, e os valores estimados da variável dependente, $\hat{y}_i$).

Coeficiente de determinação múltiplo Uma medida da eficiência de ajuste da equação de regressão múltipla estimada. Ele pode ser interpretado como a proporção da variabilidade na variável dependente que é explicada pela equação de regressão estimada.

Coeficiente de determinação múltiplo ajustado Uma medida da eficiência de ajuste da equação de regressão múltipla estimada que ajusta o número de variáveis independentes no modelo e, desse modo, evita superestimar o impacto de se acrescentar mais variáveis independentes.

Multicolinearidade O termo usado para descrever a correlação entre as variáveis independentes.

Variável qualitativa independente Uma variável independente com dados qualitativos.

Variável simulada (*Dummy variable*) Uma variável usada para modelar o efeito de variáveis qualitativas independentes. Uma variável simulada pode assumir somente os valores zero ou um.

Fórmulas-Chave

Modelo de Regressão Múltipla

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \epsilon \tag{13.1}$$

Equação de Regressão Múltipla

$$E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p \tag{13.2}$$

Equação de Regressão Múltipla Estimada

$$\hat{y} = b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_p x_p \tag{13.3}$$

Critério dos Mínimos Quadrados

$$\min \Sigma (y_i - \hat{y}_i)^2 \tag{13.4}$$

Relação Entre SST, SSR e SSE

$$\text{SST} = \text{SSR} + \text{SSE} \tag{13.7}$$

Coeficiente de Determinação Múltiplo

$$R^2 = \frac{\text{SSR}}{\text{SST}} \tag{13.8}$$

Coeficiente de Determinação Múltiplo Ajustado

$$R_a^2 = 1 - (1 - R^2)\frac{n-1}{n-p-1} \tag{13.9}$$

Regressão Média Quadrática

$$\text{MSR} = \frac{\text{SSR}}{p} \tag{13.12}$$

Erro Médio Quadrático

$$\text{MSE} = \frac{\text{SSE}}{n-p-1} \tag{13.13}$$

Estatística de Teste *F*

$$F = \frac{\text{MSR}}{\text{MSE}} \tag{13.14}$$

Estatística de Teste t

$$t = \frac{b_i}{s_{b_i}} \tag{13.15}$$

Exercícios Suplementares

39. O responsável pelas matrículas escolares (*admissions officer*) do Clearwater College desenvolveu a seguinte equação de regressão estimada relacionando o GPA acadêmico final do estudante com a pontuação SAT em matemática e o GPA obtido no curso colegial,

$$\hat{y} = 1{,}41 + 0{,}235x_1 + 0{,}00486x_2$$

em que

x_1 = *grade point average* – GPA obtido no colégio
x_2 = pontuação SAT em matemática
y = *grade point average* – GPA acadêmico final

a. Interprete os coeficientes dessa equação de regressão estimada.
b. Estime o GPA acadêmico final de um estudante que tem a média 84 no curso colegial e uma pontuação 540 no exame SAT de matemática.

40. O diretor de pessoal da Electronics Associates desenvolveu a seguinte equação de regressão estimada relacionando a pontuação que o empregado obteve em um teste de satisfação no trabalho com seu tempo de serviço e seu nível de remuneração

$$\hat{y} = 14{,}4 + 8{,}69x_1 + 13{,}5x_2$$

em que

x_1 = tempo de serviço (em anos)
x_2 = nível de remuneração (em dólares)
y = pontuação no teste de satisfação no trabalho (pontuações mais altas indicam melhor satisfação no trabalho)

a. Interprete os coeficientes dessa equação de regressão estimada.
b. Desenvolva uma estimativa da pontuação no teste de satisfação no trabalho para um empregado que tenha quatro anos de serviço e ganhe US$ 6,50 por hora.

41. Apresentamos a seguir o resultado computadorizado parcial de uma análise de regressão:

```
The regression equation is
Y = 8.103 + 7.602 X1 + 3.111 X2

Predictor          Coef        SE Coef        T
Constant          ______        2.667       _____
X1                ______        2.105       _____
X2                ______        0.613       _____

S = 3.335     R-sq = 92.3%     R-sq(adj) = _____%

Analysis of Variance

SOURCE             DF          SS          MS          F
Regression        _____       1612        _____      _____
Residual Error     12         _____       _____
Total             _____       _____
```

a. Calcule as razões t apropriadas.
b. Teste a significância de β_1 e β_2, sendo $\alpha = 0{,}05$.
c. Calcule as entradas nas colunas DF, SS e MS.
d. Calcule R_a^2.

42. Lembre-se de que no exercício 39, o responsável pelas matrículas escolares do Clearwater College desenvolveu a seguinte equação de regressão estimada relacionando o GPA acadêmico final do estudante com a pontuação SAT em matemática e o GPA obtido no curso colegial:

$$\hat{y} = 1{,}41 + 0{,}0235x_1 + 0{,}00486x_2$$

em que

x_1 = *grade point average* – GPA obtido no curso colegial
x_2 = pontuação SAT em matemática
y = *grade point average* – GPA acadêmico final

Uma parte da saída computadorizada do Minitab é apresentada a seguir:

```
The regression equation is
Y = -1.41 + .0235 X1 + .00486 X2

Predictor          Coef       SE Coef         T
Constant        -1.4053        0.4848      ______
X1             0.023467      0.008666      ______
X2               ______      0.001077      ______

S = 0.1298     R-sq = ______     R-sq(adj) = ______

Analysis of Variance

SOURCE            DF         SS          MS         F
Regression      ______    1.76209      ______     ______
Residual Error  ______    _______      ______
Total              9      1.88000
```

a. Preencha os lançamentos que faltam nessa saída de dados.
b. Calcule F e faça um teste ao nível de significância 0,05 para ver se uma relação significativa está presente.
c. A equação de regressão estimada proporcionou um bom ajuste para os dados? Explique.
d. Use o teste t e $\alpha = 0{,}05$ para testar H_0: $\beta_1 = 0$ e H_0: $\beta_2 = 0$.

43. Lembre-se de que no exercício 40, o diretor de pessoal da Electronics Associates desenvolveu a seguinte equação de regressão estimada relacionando a pontuação que o empregado obteve em um teste de satisfação no trabalho com seu tempo de serviço e seu nível de remuneração

$$\hat{y} = 14{,}4 + 8{,}69x_1 + 13{,}5x_2$$

em que

x_1 = tempo de serviço (em anos)
x_2 = nível de remuneração (em dólares)
y = pontuação no teste de satisfação no trabalho (pontuações mais altas indicam melhor satisfação no trabalho)

Uma parte da saída computadorizada do Minitab é apresentada a seguir:

```
The regression equation is
Y = 14.4 - 8.69 X1 + 13.52 X2

Predictor          Coef       SE Coef         T
Constant         14.448         8.191       1.76
X1               ______         1.555      ______
X2               13.517         2.085      ______

S = 3.773      R-sq = ______%    R-sq(adj) = ______%

Analysis of Variance

SOURCE            DF         SS          MS         F
Regression         2      _______      ______     ______
Residual Error  ______      71.17      ______
Total              7        720.0
```

a. Preencha os lançamentos que faltam nessa saída de dados.
b. Calcule F e faça um teste usando $\alpha = 0{,}05$ para ver se uma relação significativa está presente.
c. A equação de regressão estimada proporcionou um bom ajuste para os dados? Explique.
d. Use o teste t e $\alpha = 0{,}05$ para testar H_0: $\beta_1 = 0$ e H_0: $\beta_2 = 0$.

44. A revista *SmartMoney* avaliou 65 regiões metropolitanas para determinar onde os preços das casas eram mais altos. Uma cidade ideal obteria uma pontuação 100 se todos os fatores medidos fossem os mais favoráveis possíveis. Regiões com pontuações a partir de 60 são consideradas de primeira linha em termos de valorização de preço, e as regiões com uma pontuação abaixo de 50 podem sofrer deterioração de preços das moradias. Dois dos fatores avaliados foram a resistência da região à recessão econômica e sua acessibilidade e preços. Ambos os fatores foram classificados usando-se uma escala que varia de 0 (pontuação baixa) a 10 (pontuação alta). Os dados obtidos de uma amostra de 20 cidades avaliadas pela *SmartMoney* são apresentados a seguir (*SmartMoney*, fevereiro de 2002).

Região Metropolitana	Resistência à Recessão	Acessibilidade de Preços	Pontuação
Tucson	10	7	70,7
Fort Worth	10	7	68,5
San Antonio	6	8	65,5
Richmond	8	6	63,6
Indianápolis	4	8	62,5
Filadélfia	0	10	61,9
Atlanta	2	6	60,7
Phoenix	4	5	60,3
Cincinnati	2	7	57,0
Miami	6	5	56,5
Hartford	0	7	56,2
Birmingham	0	8	55,7
San Diego	8	2	54,6
Raleigh	2	7	50,9
Oklahoma City	1	6	49,6
Orange County	4	2	49,1
Denver	4	4	48,6
Los Angeles	0	7	45,7
Detroit	0	5	44,3
Nova Orleans	0	5	41,2

ARQUIVO DA INTERNET
HomeValue

a. Desenvolva uma equação de regressão estimada que possa ser usada para prever a pontuação, dada a classificação de resistência à recessão. No nível de significância 0,05, teste se há uma relação significativa.
b. A equação de regressão estimada desenvolvida no item (a) proporcionou um bom ajuste para os dados? Explique.
c. Desenvolva uma equação de regressão estimada que possa ser usada para prever a pontuação, dada a classificação de resistência à recessão e a classificação de acessibilidade de preços. No nível de significância 0,05, teste a significância global.

45. O mercado atual oferece ampla variedade de escolha para os compradores de veículos utilitários esportivos e picapes. Um fator importante para muitos compradores é o preço de revenda do veículo. A tabela a seguir apresenta o preço de revenda (%) depois de dois anos e o preço de varejo sugerido de dez utilitários esportivos, dez picapes pequenas e dez caminhonetes grandes (*Kipliger's New Cars & Trucks 2000 Buyer's Guide*).

Marca e Modelo	Tipo de Veículo	Preco de Varejo Sugerido(US$)	Valor de Revenda (%)
Chevrolet Blazer LS	Utilitário esportivo	19.495	55
Ford Explorer Sport	Utilitário esportivo	20.495	57
GMC Yukon XL 1500	Utilitário esportivo	26.789	67
Honda CR-V	Utilitário esportivo	18.965	65
Isuzu VehiCross	Utilitário esportivo	30.186	62
Jeep Cherokee Limited	Utilitário esportivo	25.745	57
Mercury Mountaineer Monterrey	Utilitário esportivo	29.895	59

ARQUIVO DA INTERNET
Trucks

Marca e Modelo	Tipo de Veículo	Preco de Varejo Sugerido(US$)	Valor de Revenda (%)
Nissan Pathfinder XE	Utilitário esportivo	26.919	54
Toyota 4Runner	Utilitário esportivo	22.418	55
Toyota RAV4	Utilitário esportivo	17.148	55
Chevrolet S-10 Extended Cab	Picape pequena	18.847	46
Dodge Dakota Club Cab Sport	Picape pequena	16.870	53
Ford Ranger XLT Regular Cab	Picape pequena	18.510	48
Ford Ranger XLT Supercab	Picape pequena	20.225	55
GMC Sonoma Regular Cab	Picape pequena	16.938	44
Isuzu Hombre Spacecab	Picape pequena	18.820	41
Mazda B4000 SE Cab Plus	Picape pequena	23.050	51
Nissan Frontier XE Regular Cab	Picape pequena	12.110	51
Toyota Tacoma Xtracab	Picape pequena	18.228	49
Toyota Tacoma Xtracab V6	Picape pequena	19.318	50
Chevrolet K2500	Picape grande	24.417	60
Chevrolet Silverado 2500 Ext	Picape grande	24.140	64
Dodge Ram 1500	Picape grande	17.460	54
Dodge Ram Quad Cab 2500	Picape grande	32.770	63
Dodge Ram Regular Cab 2500	Picape grande	23.140	59
Ford F150 XL	Picape grande	22.875	58
Ford F350 Super Duty Crew Cab XL	Picape grande	34.295	64
GMC New Sierra 1500 Ext Cab	Picape grande	27.089	68
Toyota Tundra Access Cab Limited	Picape grande	25.605	53
Toyota Tundra Regular Cab	Picape grande	15.835	58

a. Desenvolva uma equação de regressão estimada que possa ser usada para prever o valor de revenda, dado o preço de varejo sugerido. No nível de significância 0,05, teste se há uma relação significativa.
b. A equação de regressão estimada desenvolvida no item (a) proporcionou um bom ajuste para os dados? Explique.
c. Desenvolva uma equação de regressão estimada que possa ser usada para prever o valor de revenda, dado o preço de varejo sugerido e o tipo de veículo.
d. Use o teste F para determinar a significância dos resultados da regressão. No nível de significância de 0,05, qual é a sua conclusão?

46. O *Fuel Economy Guide* do U.S. Department of Energy publica dados sobre a eficiência de combustível para carros e caminhões. Parte dos dados de 35 picapes-padrão produzidas pela Chevrolet e General Motors é apresentada a seguir (http://www.fueleconomy.gov, 21 de março de 2003). A coluna intitulada Tração indica se o veículo tem tração em duas rodas (T2R) ou se tem tração nas quatro rodas (T4R). A coluna intitulada Cilindradas apresenta a capacidade em litros das cilindradas do motor, a coluna Cilindros especifica o número de cilindros que o motor tem, e a coluna intitulada Transmissão indica se o caminhão tem transmissão automática ou manual. A coluna intitulada MPG Cidade indica a avaliação da eficiência de combustível em termos de milhas por galão[6] (mpg) quando o veículo roda na cidade.

Caminhão	Nome	Tração	Cilindradas	Cilindros	Transmissão	MPG Cidade
1	C1500 Silverado	T2R	4,3	6	Automática	15
2	C1500 Silverado	T2R	4,3	6	Manual	15
3	C1500 Silverado	T2R	4,8	8	Automática	15
4	C1500 Silverado	T2R	4,8	8	Manual	16
5	C1500 Silverado	T2R	5,3	8	Automática	11
⋮	⋮	⋮	⋮	⋮	⋮	⋮
32	K1500 Sierra	T4R	5,3	8	Automática	15
33	K1500 Sierra	T4R	5,3	8	Automática	15
34	Sonoma	T4R	4,3	6	Automática	17
35	Sonoma	T4R	4,3	6	Manual	15

[6] NT: Galão – Medida de capacidade que equivale a aproximadamente 3,78 litros (Estados Unidos).

a. Desenvolva a equação de regressão estimada que possa ser usada para prever a eficiência de combustível quando o veículo roda na cidade, dado o número de cilindradas. Teste a significância usando $\alpha = 0,05$.
b. Considere o acréscimo de uma variável simulada Tração4, em que o valor de Tração4 é 0 se o caminhão tiver tração em duas rodas e 1 se o caminhão tiver tração nas quatro rodas. Desenvolva a equação de regressão estimada que possa ser usada para prever a eficiência de combustível quando se dirige na cidade, dado o número de cilindradas do motor e a variável simulada Tração4.
c. Use $\alpha = 0,05$ para determinar se a variável simulada acrescentada no item (b) é significativa.
d. Considere o acréscimo da variável simulada OitoCil, em que o valor de OitoCil é 0 se o motor do caminhão tiver seis cilindros e 1 se o motor do caminhão tiver oito cilindros. Desenvolva a equação de regressão estimada que possa ser usada para prever a eficiência de combustível quando se dirige na cidade, dado o número de cilindradas e as variáveis simuladas Tração4 e OitoCil.
e. Em relação à equação de regressão estimada desenvolvida no item (d), teste a significância global e a significância individual usando $\alpha = 0,05$.

Estudo de Caso 1 – Consumer Research, Inc.

A Consumer Research, Inc. é uma entidade independente que realiza pesquisas sobre as atitudes e comportamentos dos consumidores para uma série de empresas. Em um estudo, um cliente solicitou a investigação das características de consumo que possam ser usadas para prever o valor cobrado de usuários de cartões de crédito. Foram coletados dados sobre a renda anual, tamanho da família e gastos anuais com cartões de crédito de uma amostra de 50 consumidores. Os dados a seguir encontram-se no site, no conjunto de dados (*data set*) intitulado Consumer.

Renda (em milhares de dólares)	Tamanho da Família	Valor Cobrado (US$)	Renda (em milhares de dólares)	Tamanho da Família	Valor Cobrado (US$)
54	3	4.016	54	6	5.573
30	2	3.159	30	1	2.583
32	4	5.100	48	2	3.866
50	5	4.742	34	5	3.586
31	2	1.864	67	4	5.037
55	2	4.070	50	2	3.605
37	1	2.731	67	5	5.345
40	2	3.348	55	6	5.370
66	4	4.764	52	2	3.890
51	3	4.110	62	3	4.705
25	3	4.208	64	2	4.157
48	4	4.219	22	3	3.579
27	1	2.477	29	4	3.890
33	2	2.514	39	2	2.972
65	3	4.214	35	1	3.121
63	4	4.965	39	4	4.183
42	6	4.412	54	3	3.730
21	2	2.448	23	6	4.127
44	1	2.995	27	2	2..921
37	5	4.171	26	7	4.603
62	6	5.678	61	2	4.273
21	3	3.623	30	2	3.067
55	7	5.301	22	4	3.074
42	2	3.020	46	5	4.820
41	7	4.828	66	4	5.149

ARQUIVO DA INTERNET
Consumer

Relatório Administrativo

1. Use métodos de estatística descritiva para resumir os dados. Comente os resultados.
2. Desenvolva equações de regressão estimadas, primeiramente usando a renda anual como variável independente e depois utilizando o tamanho da família como variável independente. Qual variável prevê melhor os encargos anuais de cartões de crédito? Discuta suas conclusões.
3. Desenvolva uma equação de regressão estimada tendo a renda anual e o tamanho da família como variáveis independentes. Discuta suas conclusões.
4. Qual é o encargo anual com cartão e crédito previsto para uma família de três pessoas que tem uma renda anual de US$ 40 mil?
5. Discuta a necessidade de outras variáveis independentes que poderiam ser acrescentadas ao modelo. Quais variáveis adicionais poderiam ser úteis?

Estudo de Caso 2 – Previsão das Pontuações no Exame de Proficiência Escolar

Para prever como um distrito escolar se classificaria quando fosse levada em conta a pobreza e outras medições de renda, o *Cincinnati Enquirer* coletou dados do Education Management Services, do Ohio Department of Education e do Ohio Department of Taxation (*The Cincinnati Enquirer*, 30 de novembro de 1997). Primeiramente, o jornal obteve dados sobre o índice de aprovação em matemática, leitura, ciências, redação e nos exames de conhecimento de cidadania ministrados a alunos da quarta, sexta, nona e 12ª séries[7] no início de 1996. Combinando esses dados, eles calcularam uma porcentagem global dos estudantes de cada distrito que foram aprovados nos exames.

A porcentagem de estudantes de um distrito escolar que participam do programa Aid for Dependent Children (Auxílio para Crianças Carentes – ADC), a porcentagem dos que têm direito a merendas gratuitas ou a preços reduzidos, e a mediana da renda familiar no distrito escolar também foram registradas. Parte dos dados coletados relativos aos 608 distritos escolares é apresentada a seguir. O conjunto de dados completo está disponível no site www.thomsonlearning.com.br/estatapl.htm, no arquivo intitulado Enquirer.

ARQUIVO DA INTERNET

Enquirer

Classificação	Distrito Escolar	Município	% dos Aprovados	% no ADC	% Merenda Gratuita	Mediana da Renda (US$)
1	Ottawa Hills Local	Lucas	93,85	0,11	0,00	48.231
2	Wyoming City	Hamilton	93,08	2,95	4,59	42.672
3	Oakwood City	Montgomery	92,92	0,20	0,38	42.403
4	Madeira City	Hamilton	92,37	1,50	4,83	32.889
5	Indian Hill Ex Vill	Hamilton	91,77	1,23	2,70	44.135
6	Solon City	Cuyahoga	90,77	0,68	2,24	34.993
7	Chagrin Falls Ex Vill	Cuyahoga	89,89	0,47	0,44	38.921
8	Mariemont City	Hamilton	89,80	3,00	2,97	31.823
9	Upper Arlington City	Franklin	89,77	0,24	0,92	38.358
10	Granville Ex Vill	Licking	89,22	1,14	0,00	36.235
⋮	⋮	⋮	⋮	⋮	⋮	⋮

Os dados foram classificados com base nos valores da coluna intitulada Porcentagem dos Aprovados; esses dados são a porcentagem global dos estudantes que foram aprovados nos exames. Os dados na coluna intitulada Porcentagem no ADC são a porcentagem dos estudantes de cada distrito escolar que fazem parte do programa ADC, e os dados na coluna intitulada Porcentagem Merenda Gratuita são a porcentagem de estudantes que se habilitam a receber merendas gratuitas ou a preços reduzidos. A coluna intitulada Mediana da Renda indica a mediana da renda familiar de cada distrito escolar. Em relação a cada distrito escolar, também é indicado em qual município ele se encontra. Observe que em alguns casos o valor

[7] NT: Nos Estados Unidos, há a *elementary school*, os seis primeiros anos de estudo, em que o aluno aprende as matérias básicas. Depois, ele passa à *junior high-school*, escola intermediária, que geralmente inclui a 7ª, 8ª e 9ª séries. Finalmente, o aluno cursa a *senior high-school*, que oferece os últimos anos da educação secundária; geralmente, a 10ª, 11ª e 12ª séries.

inserido na coluna Porcentagem Merenda Gratuita é 0, indicando que o distrito não participava do programa de merenda gratuita.

Relatório Administrativo

Use os métodos apresentados neste capítulo e nos anteriores para analisar esse conjunto de dados. Apresente um resumo de sua análise, incluindo os resultados estatísticos, conclusões e recomendações fundamentais no formato de relatório administrativo. Inclua quaisquer materiais técnicos que julgar necessários em um apêndice.

Estudo de Caso 3 – Doações de Ex-Alunos

As doações de ex-alunos são uma fonte importante de receitas para colégios e universidades. Se os administradores pudessem determinar os fatores que influem no aumento da porcentagem de ex-alunos que fazem doações, talvez pudessem implementar políticas que levassem a um aumento das receitas. Pesquisas mostram que os estudantes que estão mais satisfeitos em seus contatos com os professores têm mais probabilidade de graduar-se. Em conseqüência, poder-se-ia imaginar que classes menores e uma razão menor entre professores e alunos poderiam acarretar maior porcentagem de graduados satisfeitos, o que, por sua vez, poderia levar a um aumento na porcentagem de ex-alunos que fazem doações. A Tabela 13.7 apresenta dados de 48 universidades federais (*America's Best Colleges*, 2000). A coluna intitulada Índice de Graduação é a porcentagem de estudantes que inicialmente se matricularam na universidade e se diplomaram. A coluna intitulada Porcentagem de Classes com Menos de 20 exibe a porcentagem de classes disponíveis com menos de 20 alunos. A coluna intitulada Razão Estudantes/Professor refere-se ao número de estudantes matriculados dividido pelo número total de professores. Finalmente, a coluna intitulada Índice de Doação de Ex-alunos é a porcentagem de ex-alunos que fizeram doações à universidade.

Relatório Administrativo

1. Use métodos de estatística descritiva para resumir os dados.
2. Desenvolva uma equação de regressão estimada que possa ser usada para prever o índice de doações de ex-alunos, dado o número de estudantes que se graduam. Discuta suas conclusões.
3. Usando os dados apresentados, desenvolva uma equação de regressão estimada que possa ser usada para prever o índice de doações feitas por ex-alunos.
4. Quais conclusões e recomendações você é capaz de deduzir de sua análise?

Apêndice 13.1 – Regressão Múltipla com o Minitab

Na Seção 13.2, discutimos a solução computadorizada de problemas de regressão múltipla ao apresentarmos a saída de dados do Minitab correspondente ao problema da Butler Trucking Company. Neste apêndice, descrevemos as etapas necessárias para gerar a solução computadorizada do Minitab. Primeiramente, os dados devem ser inseridos em uma planilha do Minitab. As milhas percorridas são inseridas na coluna C1, o número de entregas é inserido na coluna C2 e os tempos de viagem (em horas) são inseridos na coluna C3. Os nomes das variáveis, Miles (Milhas), Deliv (Entrega) e Time (Tempo) foram inseridos como cabeçalhos de coluna na planilha. Nas etapas subseqüentes, nos referirmos aos dados usando os nomes das variáveis Miles, Deliv e Time ou os indicadores de coluna C1, C2 e C3. As etapas a seguir descrevem como usar o Minitab para produzir os resultados de regressão apresentados na Figura 13.4.

ARQUIVO DA INTERNET
Butler

Etapa 1. Selecione o menu **Stat**
Etapa 2. Selecione o menu **Regression**
Etapa 3. Escolha a opção **Regression**
Etapa 4. Quando a caixa de diálogo **Regression** aparecer:
Digite Time (Tempo) na caixa **Response**
Digite Miles (Milhas) e Deliv (Entrega) na caixa **Predictors**
Dê um clique em **OK**

ARQUIVO DA INTERNET **Alumni**

Tabela 13.7 Dados de 48 Universidades Federais

Universidade	Estado	Índice de Graduação	% de Classes com Menos de 20	Razão Estudantes/ Professor	Índice de Doação de Ex-alunos
Boston College	MA	85	39	13	25
Brandeis University	MA	79	68	8	33
Brown University	RI	93	60	8	40
California Institute of Technology	CA	85	65	3	46
Carnegie Mellon University	PA	75	67	10	28
Case Western Reserve Univ.	OH	72	52	8	31
College of William and Mary	VA	89	45	12	27
Columbia University	NY	90	69	7	31
Cornell University	NY	91	72	13	35
Dartmouth College	NH	94	61	10	53
Duke University	NC	92	68	8	45
Emory University	GA	84	65	7	37
Georgetown University	PA	91	54	10	29
Harvard University	MA	97	73	8	46
Johns Hopkins University	MD	89	64	9	27
Lehigh University	PA	81	55	11	40
Massachusetts Inst. of Technology	MA	92	65	6	44
New York University	NY	72	63	13	13
Northwestern University	IL	90	66	8	30
Pennsylvania State Univ.	PA	80	32	19	21
Princeton University	NJ	95	68	5	67
Rice University	TX	92	62	8	40
Stanford University	CA	92	69	7	34
Tufts University	MA	87	67	9	29
Tulane University	LA	72	56	12	17
U. of California–Berkeley	CA	83	58	17	18
U. of California–Davis	CA	74	32	19	7
U. of California–Irvine	CA	74	42	20	9
U. of California–Los Angeles	CA	78	41	18	13
U. of California–San Diego	CA	80	48	19	8
U. of California–Santa Barbara	CA	70	45	20	12
U. of Chicago	IL	84	65	4	36
U. of Florida	FL	67	31	23	19
U. of Illinois–Urbana Champaign	IL	77	29	15	23
U. of Michigan–Ann Arbor	MI	83	51	15	13
U. of North Carolina–Chapel Hill	NC	82	40	16	26
U. of Notre Dame	IN	94	53	13	49
U. of Pennsylvania	PA	90	65	7	41
U. of Rochester	NY	76	63	10	23
U. of Southern California	CA	70	53	13	22
U. of Texas–Austin	TX	66	39	21	13
U. of Virginia	VA	92	44	13	28
U. of Washington	WA	70	37	12	12
U. of Wisconsin–Madison	WI	73	37	13	13
Vanderbilt University	TN	82	68	9	31
Wake Forest University	NC	82	59	11	38
Washington University–St. Louis	MO	86	73	7	33
Yale University	CT	94	77	7	50

Apêndice 13.2 – Regressão Múltipla com o Excel

Na Seção 13.2, discutimos a solução computadorizada de problemas de regressão múltipla ao apresentarmos a saída de dados do Minitab correspondente ao problema da Butler Trucking Company. Neste apên-

dice, descrevemos como usar a ferramenta Regressão do Excel para desenvolvermos a equação de regressão múltipla estimada do problema da Butler Trucking Company. Consulte a Figura 13.10 à medida que descrevermos as tarefas envolvidas. Primeiramente, os rótulos Tarefas, Milhas, Entregas e Tempo são inseridos nas células A1:D1 da planilha e os dados amostrais nas células B2:D11. Os números 1 a 10 nas células A2:D1 identificam cada observação.

ARQUIVO DA INTERNET
Butler

As etapas a seguir descrevem como usar a ferramenta Regressão na análise de regressão múltipla.

Etapa 1. Selecione o menu **Ferramentas**
Etapa 2. Escolha a opção **Análise de Dados**
Etapa 3. Escolha **Regressão** na lista de Ferramentas de Análise

Figura 13.10 Saída de dados do Excel para o problema da Butler Trucking com duas variáveis independentes

	A	B	C	D	E	F	G	H	I
1	Tarefa	Milhas	Entregas	Tempo					
2	1	100	4	9.3					
3	2	50	3	4.8					
4	3	100	4	8.9					
5	4	100	2	6.5					
6	5	50	2	4.2					
7	6	80	2	6.2					
8	7	75	3	7.4					
9	8	65	4	6					
10	9	90	3	7.6					
11	10	90	2	6.1					
12									
13	RESUMO DAS SAÍDAS								
14									
15	*Estatística de Regressão*								
16	R Múltipla	0.9507							
17	R-Sq	0.9038							
18	R-Sq Ajustado	0.8763							
19	Erro Padrão	0.5731							
20	Observações	10							
21									
22	ANOVA								
23		*gl*	*SQ*	*MQ*	*F*	*Significância F*			
24	Regressão	2	21.6006	10.8003	32.8784	0.0003			
25	Resíduo	7	2.2994	0.3285					
26	Total	9	23.9						
27									
28		*Coeficientes*	*Erro Padrão*	*Estat. t*	*Valor p*	*Mín. 95%*	*Máx. 95%*	*Mín. 99.0%*	*Máx. 99.0%*
29	Intercepção	-0.8687	0.9515	-0.9129	0.3916	-3.1188	1.3813	-4.1986	2.4612
30	Milhas	0.0611	0.0099	6.1824	0.0005	0.0378	0.0845	0.0265	0.0957
31	Entregas	0.9234	0.2211	4.1763	0.0042	0.4006	1.4463	0.1496	1.6972

Etapa 4. Quando a caixa de diálogo Regressão aparecer
Digite D1:D11 na caixa **Intervalo Y de Entrada**
Digite B1:C11 na caixa **Intervalo X de Entrada**
Marque a opção **Rótulos**
Marque a opção **Nível de Confiança**
Digite 99 na caixa **Nível de Confiança**
Marque a opção **Intervalo de Saída**
Digite A13 na caixa **Intervalo de Saída** (para identificar o canto superior esquerdo da parte da planilha em que a saída aparecerá)
Dê um clique em **OK**

Na saída do Excel apresentada na Figura 13.10, o rótulo da variável independente x_1 é Milhas (veja a célula A30), e o rótulo da variável independente x_2 é Entregas (veja a célula A31). A equação de regressão estimada é:

$$\hat{y} = -0{,}8687 + 0{,}611x_1 + 0{,}9234x_2$$

Note que usar a ferramenta Regressão do Excel para regressão múltipla é quase o mesmo que usá-la para regressão linear simples. A principal diferença é que no caso da regressão múltipla é necessário um intervalo maior de células para identificar as variáveis independentes.

APÊNDICE A

Referências e Bibliografia

Geral

Bowerman, B. L., e R. T. O'Connell, *Applied Statistics: Improving Business Processes*, Irwin, 1996.
Freedman, D., R. Pisani e R. Purves, *Statistics*, 3. ed., W. W. Norton, 1997.
Freund, J. E., *Mathematical Statistics*, 5. ed., Prentice-Hall, 1992.
Hogg, R. V. e A. T. Craig, *Introduction to Mathematical Statistics*, 5. ed., Prentice-Hall, 1995.
McClave, J. T. e P. G. Benson, *Statistics: For Business and Economics*, 6. ed., MacMillan, 1994.
Moore, D. S. e G. P. McCabe, *Introduction to the Practice of Statistics*, 3. ed., W. H. Freeman & Co.,1998.
Neter, J., W. Wasserman e G. A. Whitmore, *Applied Statistics*, 4. ed., Allyn & Bacon, 1992.
Roberts, H., *Data Analysis for Managers with Minitab*, Scientific Press, 1991.
Ryan, B. F. e B. L. Joiner, *Minitab Handbook*, 3. ed., Duxbury Press, 1994.
Tanur, J. M., *Statistics: A Guide to the Unknown*, 3. ed., Brooks/ Cole, 1989.
Tukey, J. W., *Exploratory Data Analysis*, Addison-Wesley, 1977.

Probabilidade

Feller, W., An *Introduction to Probability Theory and Its Application*, vol. I, 3. ed., Wiley, 1968.
Hogg, R. V. e E. A. Tanis, *Probability and Statistical Inference*, 5. ed., Prentice-Hall, 1996.
Ross, S. M., *Introduction to Probability Models*, 6. ed., Academic Press, 1997.
Wackerly, D. D., W. Mendenhall e R. L. Scheaffer, *Mathematical Statistics with Applications*, 5. ed., PWS, 1996.

Análise de Regressão

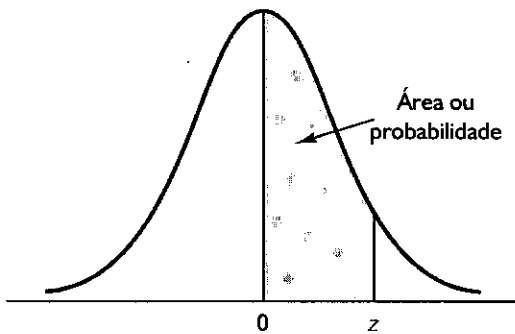
Belsley, D. A., *Conditioning Diagnostics: Collinearity and Weak Data in Regression*, Wiley, 1991.
Chatterjee, S. e B. Price, *Regression Analysis by Example*, 2. ed., Wiley, 1991.
Cook, R. D. e S. Weisberg, *Residuals and Influence in Regression*, Chapman & Hall, 1982.
Draper, N. R., S. Draper e H. Smith, *Applied Regression Analysis*, 3. ed., Wiley, 1998.
Graybill, F. A. e H. K. Iyer, *Regression Analysis: Concepts and Applications*, Duxbury Press, 1994.
Kleinbaum, D. G., L. L. Kupper e K. E. Muller, *Applied Regression Analysis and Other Multivariate Methods*, 3. ed., Duxbury Press, 1997.

Myers, R. H., *Classical and Modern Regression with Applications*, 2. ed., PWS, 1990.
Neter, J., W. Wasserman e M. H. Kutner, *Applied Linear Statistical Models: Regression, Analysis of Variance and Experimental Design*, 4. ed., Irwin, 1990.
Wonnacott, T. H. e R. J. Wonnacott, *Regression: A Second Course in Statistics*, Krieger, 1986.

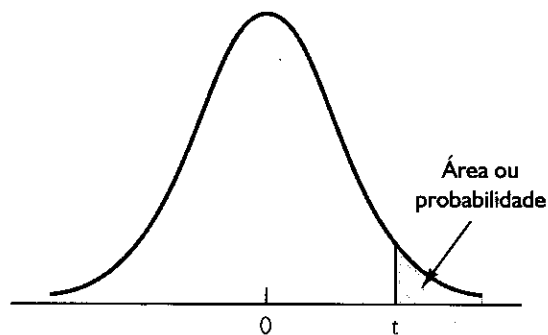
Controle de Qualidade

Deming, W. E., *Quality, Productivity, and Competitive Position*, MIT, 1982.
Duncan, A. J., *Quality Control and Industrial Statistics*, 5. ed., Irwin, 1986.
Evans, J. R. e W. M. Lindsay, *The Management and Control of Quality*, 3. ed., West/ Wadsworth, 1995.
Hunt, V. D. e H. S. Gitlow, *Managing for Quality: Integrating Quality and Business Strategy*, Irwin, 1992.
Ishikawa, K., *Introduction to Quality Control*, Quality Resources, 1990.
Juran, J. M. e F. M. Gryna, *Quality Planning and Analysis: From Product Development Through Use*, 3. ed., McGraw- Hill, 1993.
Montgomery, D. C., *Introduction to Statistical Quality Control*, 3. ed., Wiley, 1996.

Tabelas

Tabela 1 Distribuição Normal-Padrão

Os registros na tabela fornecem a área abaixo da curva entre a média e z desvios padrão acima da média. Por exemplo, para $z = 1,25$ a área abaixo da curva entre a média e z é 0,3944.

z	0,00	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
0,0	0,0000	0,0040	0,0080	0,0120	0,0160	0,0199	0,0239	0,0279	0,0319	0,0359
0,1	0,0398	0,0438	0,0478	0,0517	0,0557	0,0596	0,0636	0,0675	0,0714	0,0753
0,2	0,0793	0,0832	0,0871	0,0910	0,0948	0,0987	0,1026	0,1064	0,1103	0,1141
0,3	0,1179	0,1217	0,1255	0,1293	0,1331	0,1368	0,1406	0,1443	0,1480	0,1517
0,4	0,1554	0,1591	0,1628	0,1664	0,1700	0,1736	0,1772	0,1808	0,1844	0,1879
0,5	0,1915	0,1950	0,1985	0,2019	0,2054	0,2088	0,2123	0,2157	0,2190	0,2224
0,6	0,2257	0,2291	0,2324	0,2357	0,2389	0,2422	0,2454	0,2486	0,2517	0,2549
0,7	0,2580	0,2611	0,2642	0,2673	0,2704	0,2734	0,2764	0,2794	0,2823	0,2852
0,8	0,2881	0,2910	0,2939	0,2967	0,2995	0,3023	0,3051	0,3078	0,3106	0,3133
0,9	0,3159	0,3186	0,3212	0,3238	0,3264	0,3289	0,3315	0,3340	0,3365	0,3389
1,0	0,3413	0,3438	0,3461	0,3485	0,3508	0,3531	0,3554	0,3577	0,3599	0,3621
1,1	0,3643	0,3665	0,3686	0,3708	0,3729	0,3749	0,3770	0,3790	0,3810	0,3830
1,2	0,3849	0,3869	0,3888	0,3907	0,3925	0,3944	0,3962	0,3980	0,3997	0,4015
1,3	0,4032	0,4049	0,4066	0,4082	0,4099	0,4115	0,4131	0,4147	0,4162	0,4177
1,4	0,4192	0,4207	0,4222	0,4236	0,4251	0,4265	0,4279	0,4292	0,4306	0,4319
1,5	0,4332	0,4345	0,4357	0,4370	0,4382	0,4394	0,4406	0,4418	0,4429	0,4441
1,6	0,4452	0,4463	0,4474	0,4484	0,4495	0,4505	0,4515	0,4525	0,4535	0,4545
1,7	0,4554	0,4564	0,4573	0,4582	0,4591	0,4599	0,4608	0,4616	0,4625	0,4633
1,8	0,4641	0,4649	0,4656	0,4664	0,4671	0,4678	0,4686	0,4693	0,4699	0,4706
1,9	0,4713	0,4719	0,4726	0,4732	0,4738	0,4744	0,4750	0,4756	0,4761	0,4767
2,0	0,4772	0,4778	0,4783	0,4788	0,4793	0,4798	0,4803	0,4808	0,4812	0,4817
2,1	0,4821	0,4826	0,4830	0,4834	0,4838	0,4842	0,4846	0,4850	0,4854	0,4857
2,2	0,4861	0,4864	0,4868	0,4871	0,4875	0,4878	0,4881	0,4884	0,4887	0,4890
2,3	0,4893	0,4896	0,4898	0,4901	0,4904	0,4906	0,4909	0,4911	0,4913	0,4916
2,4	0,4918	0,4920	0,4922	0,4925	0,4927	0,4929	0,4931	0,4932	0,4934	0,4936
2,5	0,4938	0,4940	0,4941	0,4943	0,4945	0,4946	0,4948	0,4949	0,4951	0,4952
2,6	0,4953	0,4955	0,4956	0,4957	0,4959	0,4960	0,4961	0,4962	0,4963	0,4964
2,7	0,4965	0,4966	0,4967	0,4968	0,4969	0,4970	0,4971	0,4972	0,4973	0,4974
2,8	0,4974	0,4975	0,4976	0,4977	0,4977	0,4978	0,4979	0,4979	0,4980	0,4981
2,9	0,4981	0,4982	0,4982	0,4983	0,4984	0,4984	0,4985	0,4985	0,4986	0,4986
3,0	0,4987	0,4987	0,4987	0,4988	0,4988	0,4989	0,4989	0,4989	0,4990	0,4990

Tabela 2 Distribuição *t*

Os registros na tabela fornecem valores de *t* para uma área ou probabilidade na extremidade superior da distribuição *t*. Por exemplo, com 10 graus de liberdade e uma área 0,05 na cauda superior, $t_{0,05} = 1,812$.

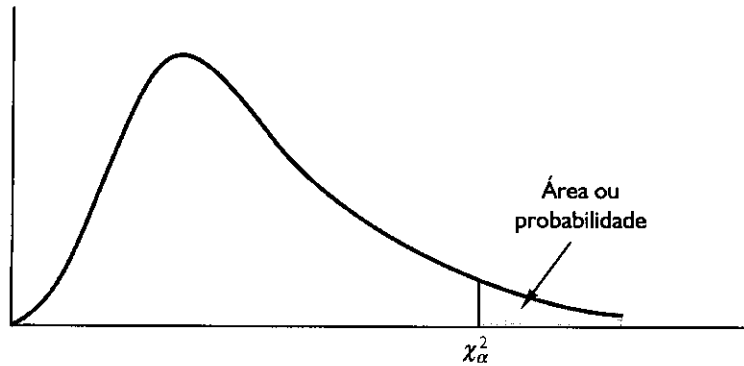
Graus de Liberdade	Área da Cauda Superior					
	0,20	0,10	0,05	0,025	0,01	0,005
1	1,376	3,078	6,314	12,706	31,821	63,656
2	1,061	1,886	2,920	4,303	6,965	9,925
3	0,978	1,638	2,353	3,182	4,541	5,841
4	0,941	1,533	2,132	2,776	3,747	4,604
5	0,920	1,476	2,015	2,571	3,365	4,032
6	0,906	1,440	1,943	2,447	3,143	3,707
7	0,896	1,415	1,895	2,365	2,998	3,499
8	0,889	1,397	1,860	2,306	2,896	3,355
9	0,883	1,383	1,833	2,262	2,821	3,250
10	0,879	1,372	1,812	2,228	2,764	3,169
11	0,876	1,363	1,796	2,201	2,718	3,106
12	0,873	1,356	1,782	2,179	2,681	3,055
13	0,870	1,350	1,771	2,160	2,650	3,012
14	0,868	1,345	1,761	2,145	2,624	2,977
15	0,866	1,341	1,753	2,131	2,602	2,947
16	0,865	1,337	1,746	2,120	2,583	2,921
17	0,863	1,333	1,740	2,110	2,567	2,898
18	0,862	1,330	1,734	2,101	2,552	2,878
19	0,861	1,328	1,729	2,093	2,539	2,861
20	0,860	1,325	1,725	2,086	2,528	2,845
21	0,859	1,323	1,721	2,080	2,518	2,831
22	0,858	1,321	1,717	2,074	2,508	2,819
23	0,858	1,319	1,714	2,069	2,500	2,807
24	0,857	1,318	1,711	2,064	2,492	2,797
25	0,856	1,316	1,708	2,060	2,485	2,787
26	0,856	1,315	1,706	2,056	2,479	2,779
27	0,855	1,314	1,703	2,052	2,473	2,771
28	0,855	1,313	1,701	2,048	2,467	2,763
29	0,854	1,311	1,699	2,045	2,462	2,756
30	0,854	1,310	1,697	2,042	2,457	2,750
31	0,853	1,309	1,696	2,040	2,453	2,744
32	0,853	1,309	1,694	2,037	2,449	2,738
33	0,853	1,308	1,692	2,035	2,445	2,733
34	0,852	1,307	1,691	2,032	2,441	2,728

Tabela 2 Distribuição *t* (continuação)

Graus de Liberdade	Área da Cauda Superior					
	0,20	0,10	0,05	0,025	0,01	0,005
35	0,852	1,306	1,690	2,030	2,438	2,724
36	0,852	1,306	1,688	2,028	2,434	2,719
37	0,851	1,305	1,687	2,026	2,431	2,715
38	0,851	1,304	1,686	2,024	2,429	2,712
39	0,851	1,304	1,685	2,023	2,426	2,708
40	0,851	1,303	1,684	2,021	2,423	2,704
41	0,850	1,303	1,683	2,020	2,421	2,701
42	0,850	1,302	1,682	2,018	2,418	2,698
43	0,850	1,302	1,681	2,017	2,416	2,695
44	0,850	1,301	1,680	2,015	2,414	2,692
45	0,850	1,301	1,679	2,014	2,412	2,690
46	0,850	1,300	1,679	2,013	2,410	2,687
47	0,849	1,300	1,678	2,012	2,408	2,685
48	0,849	1,299	1,677	2,011	2,407	2,682
49	0,849	1,299	1,677	2,010	2,405	2,680
50	0,849	1,299	1,676	2,009	2,403	2,678
51	0,849	1,298	1,675	2,008	2,402	2,676
52	0,849	1,298	1,675	2,007	2,400	2,674
53	0,848	1,298	1,674	2,006	2,399	2,672
54	0,848	1,297	1,674	2,005	2,397	2,670
55	0,848	1,297	1,673	2,004	2,396	2,668
56	0,848	1,297	1,673	2,003	2,395	2,667
57	0,848	1,297	1,672	2,002	2,394	2,665
58	0,848	1,296	1,672	2,002	2,392	2,663
59	0,848	1,296	1,671	2,001	2,391	2,662
60	0,848	1,296	1,671	2,000	2,390	2,660
61	0,848	1,296	1,670	2,000	2,389	2,659
62	0,847	1,295	1,670	1,999	2,388	2,657
63	0,847	1,295	1,669	1,998	2,387	2,656
64	0,847	1,295	1,669	1,998	2,386	2,655
65	0,847	1,295	1,669	1,997	2,385	2,654
66	0,847	1,295	1,668	1,997	2,384	2,652
67	0,847	1,294	1,668	1,996	2,383	2,651
68	0,847	1,294	1,668	1,995	2,382	2,650
69	0,847	1,294	1,667	1,995	2,382	2,649
70	0,847	1,294	1,667	1,994	2,381	2,648
71	0,847	1,294	1,667	1,994	2,380	2,647
72	0,847	1,293	1,666	1,993	2,379	2,646
73	0,847	1,293	1,666	1,993	2,379	2,645
74	0,847	1,293	1,666	1,993	2,378	2,644
75	0,846	1,293	1,665	1,992	2,377	2,643
76	0,846	1,293	1,665	1,992	2,376	2,642
77	0,846	1,293	1,665	1,991	2,376	2,641
78	0,846	1,292	1,665	1,991	2,375	2,640
79	0,846	1,292	1,664	1,990	2,374	2,639
80	0,846	1,292	1,664	1,990	2,374	2,639
81	0,846	1,292	1,664	1,990	2,373	2,638
82	0,846	1,292	1,664	1,989	2,373	2,637
83	0,846	1,292	1,663	1,989	2,372	2,636
84	0,846	1,292	1,663	1,989	2,372	2,636

Tabela 2 Distribuição *t* (continuação)

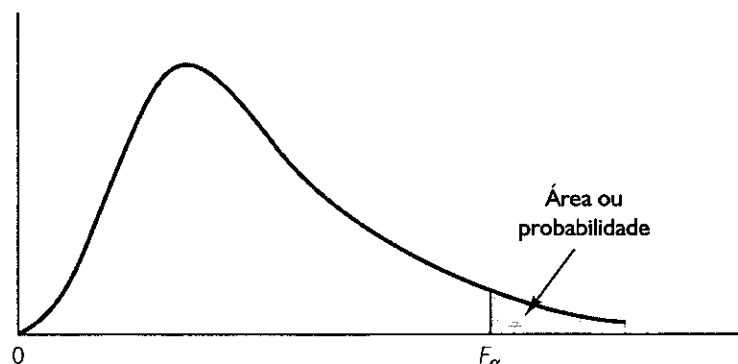
Graus de Liberdade	Área da Cauda Superior					
	0,20	0,10	0,05	0,025	0,01	0,005
85	0,846	1,292	1,663	1,988	2,371	2,635
86	0,846	1,291	1,663	1,988	2,370	2,634
87	0,846	1,291	1,663	1,988	2,370	2,634
88	0,846	1,291	1,662	1,987	2,369	2,633
89	0,846	1,291	1,662	1,987	2,369	2,632
90	0,846	1,291	1,662	1,987	2,368	2,632
91	0,846	1,291	1,662	1,986	2,368	2,631
92	0,846	1,291	1,662	1,986	2,368	2,630
93	0,846	1,291	1,661	1,986	2,367	2,630
94	0,845	1,291	1,661	1,986	2,367	2,629
95	0,845	1,291	1,661	1,985	2,366	2,629
96	0,845	1,290	1,661	1,985	2,366	2,628
97	0,845	1,290	1,661	1,985	2,365	2,627
98	0,845	1,290	1,661	1,984	2,365	2,627
99	0,845	1,290	1,660	1,984	2,364	2,626
100	0,845	1,290	1,660	1,984	2,364	2,626
∞	0,842	1,282	1,645	1,960	2,326	2,576

Tabela 3 Distribuição do Quiquadrado

Os registros na tabela fornecem valores de χ^2_α, em que A é a área ou probabilidade na cauda superior da distribuição de quiquadrado. Por exemplo, com 10 graus de liberdade e uma área de 0,01 na cauda superior, $\chi^2_{0,01} = 23,2093$.

Graus de Liberdade	Área da Cauda Superior									
	0,995	0,99	0,975	0,95	0,90	0,10	0,05	0,025	0,01	0,005
1	0,000	0,000	0,001	0,004	0,016	2,706	3,841	5,024	6,635	7,879
2	0,010	0,020	0,051	0,103	0,211	4,605	5,991	7,378	9,210	10,597
3	0,072	0,115	0,216	0,352	0,584	6,251	7,815	9,348	11,345	12,838
4	0,207	0,297	0,484	0,711	1,064	7,779	9,488	11,143	13,277	14,860
5	0,412	0,554	0,831	1,145	1,610	9,236	11,070	12,832	15,086	16,750
6	0,676	0,872	1,237	1,635	2,204	10,645	12,592	14,449	16,812	18,548
7	0,989	1,239	1,690	2,167	2,833	12,017	14,067	16,013	18,475	20,278
8	1,344	1,647	2,180	2,733	3,490	13,362	15,507	17,535	20,090	21,955
9	1,735	2,088	2,700	3,325	4,168	14,684	16,919	19,023	21,666	23,589
10	2,156	2,558	3,247	3,940	4,865	15,987	18,307	20,483	23,209	25,188
11	2,603	3,053	3,816	4,575	5,578	17,275	19,675	21,920	24,725	26,757
12	3,074	3,571	4,404	5,226	6,304	18,549	21,026	23,337	26,217	28,300
13	3,565	4,107	5,009	5,892	7,041	19,812	22,362	24,736	27,688	29,819
14	4,075	4,660	5,629	6,571	7,790	21,064	23,685	26,119	29,141	31,319
15	4,601	5,229	6,262	7,261	8,547	22,307	24,996	27,488	30,578	32,801
16	5,142	5,812	6,908	7,962	9,312	23,542	26,296	28,845	32,000	34,267
17	5,697	6,408	7,564	8,672	10,085	24,769	27,587	30,191	33,409	35,718
18	6,265	7,015	8,231	9,390	10,865	25,989	28,869	31,526	34,805	37,156
19	6,844	7,633	8,907	10,117	11,651	27,204	30,144	32,852	36,191	38,582
20	7,434	8,260	9,591	10,851	12,443	28,412	31,410	34,170	37,566	39,997
21	8,034	8,897	10,283	11,591	13,240	29,615	32,671	35,479	38,932	41,401
22	8,643	9,542	10,982	12,338	14,041	30,813	33,924	36,781	40,289	42,796
23	9,260	10,196	11,689	13,091	14,848	32,007	35,172	38,076	41,638	44,181
24	9,886	10,856	12,401	13,848	15,659	33,196	36,415	39,364	42,980	45,558
25	10,520	11,524	13,120	14,611	16,473	34,382	37,652	40,646	44,314	46,928
26	11,160	12,198	13,844	15,379	17,292	35,563	38,885	41,923	45,642	48,290
27	11,808	12,878	14,573	16,151	18,114	36,741	40,113	43,195	46,963	49,645
28	12,461	13,565	15,308	16,928	18,939	37,916	41,337	44,461	48,278	50,994
29	13,121	14,256	16,047	17,708	19,768	39,087	42,557	45,722	49,588	52,335

Tabela 3 Distribuição do Quiquadrado (continuação)

Graus de Liberdade	Área da Cauda Superior									
	0,995	0,99	0,975	0,95	0,90	0,10	0,05	0,025	0,01	0,005
30	13,787	14,953	16,791	18,493	20,599	40,256	43,773	46,979	50,892	53,672
35	17,192	18,509	20,569	22,465	24,797	46,059	49,802	53,203	57,342	60,275
40	20,707	22,164	24,433	26,509	29,051	51,805	55,758	59,342	63,691	66,766
45	24,311	25,901	28,366	30,612	33,350	57,505	61,656	65,410	69,957	73,166
50	27,991	29,707	32,357	34,764	37,689	63,167	67,505	71,420	76,154	79,490
55	31,735	33,571	36,398	38,958	42,060	68,796	73,311	77,380	82,292	85,749
60	35,534	37,485	40,482	43,188	46,459	74,397	79,082	83,298	88,379	91,952
65	39,383	41,444	44,603	47,450	50,883	79,973	84,821	89,177	94,422	98,105
70	43,275	45,442	48,758	51,739	55,329	85,527	90,531	95,023	100,425	104,215
75	47,206	49,475	52,942	56,054	59,795	91,061	96,217	100,839	106,393	110,285
80	51,172	53,540	57,153	60,391	64,278	96,578	101,879	106,629	112,329	116,321
85	55,170	57,634	61,389	64,749	68,777	102,079	107,522	112,393	118,236	122,324
90	59,196	61,754	65,647	69,126	73,291	107,565	113,145	118,136	124,116	128,299
95	63,250	65,898	69,925	73,520	77,818	113,038	118,752	123,858	129,973	134,247
100	67,328	70,065	74,222	77,929	82,358	118,498	124,342	129,561	135,807	140,170

Tabela 4 Distribuição *F*

Os registros na tabela fornecem valores de F_α, em que α é a área ou probabilidade na cauda superior da distribuição *F*. Por exemplo, com 4 graus de liberdade do numerador, 8 graus de liberdade do denominador, e uma área de 0,05 na cauda superior, $F_{0,05} = 3,48$.

Graus de Liberdade do Denominador	Área na Cauda Superior	Graus de Liberdade do Numerador																	
		1	2	3	4	5	6	7	8	9	10	15	20	25	30	40	60	100	1.000
1	0,10	39,86	49,50	53,59	55,83	57,24	58,20	58,91	59,44	59,86	60,19	61,22	61,74	62,05	62,26	62,53	62,79	63,01	63,30
	0,05	161,45	199,50	215,71	224,58	230,16	233,99	236,77	238,88	240,54	241,88	245,95	248,02	249,26	250,10	251,14	252,20	253,04	254,19
	0,025	647,79	799,48	864,15	899,60	921,83	937,11	948,20	956,64	963,28	968,63	984,87	993,08	998,09	1.001,40	1.005,60	1.009,79	1.013,16	1.017,76
	0,01	4.052,18	4.999,34	5.403,53	5.624,26	5.763,96	5.858,95	5.928,33	5.980,95	6.022,40	6.055,93	6.156,97	6.208,66	6.239,86	6.260,35	6.286,43	6.312,97	6.333,92	6.362,80
2	0,10	8,53	9,00	9,16	9,24	9,29	9,33	9,35	9,37	9,38	9,39	9,42	9,44	9,45	9,46	9,47	9,47	9,48	9,49
	0,05	18,51	19,00	19,16	19,25	19,30	19,33	19,35	19,37	19,38	19,40	19,43	19,45	19,46	19,46	19,47	19,48	19,49	19,49
	0,025	38,51	39,00	39,17	39,25	39,30	39,33	39,36	39,37	39,39	39,40	39,43	39,45	39,46	39,46	39,47	39,48	39,49	39,50
	0,01	98,50	99,00	99,16	99,25	99,30	99,33	99,36	99,38	99,39	99,40	99,43	99,45	99,46	99,47	99,48	99,48	99,49	99,50
3	0,10	5,54	5,46	5,39	5,34	5,31	5,28	5,27	5,25	5,24	5,23	5,20	5,18	5,17	5,17	5,16	5,15	5,14	5,13
	0,05	10,13	9,55	9,28	9,12	9,01	8,94	8,89	8,85	8,81	8,79	8,70	8,66	8,63	8,62	8,59	8,57	8,55	8,53
	0,025	17,44	16,04	15,44	15,10	14,88	14,73	14,62	14,54	14,47	14,42	14,25	14,17	14,12	14,08	14,04	13,99	13,96	13,91
	0,01	34,12	30,82	29,46	28,71	28,24	27,91	27,67	27,49	27,34	27,23	26,87	26,69	26,58	26,50	26,41	26,32	26,24	26,14
4	0,10	4,54	4,32	4,19	4,11	4,05	4,01	3,98	3,95	3,94	3,92	3,87	3,84	3,83	3,82	3,80	3,79	3,78	3,76
	0,05	7,71	6,94	6,59	6,39	6,26	6,16	6,09	6,04	6,00	5,96	5,86	5,80	5,77	5,75	5,72	5,69	5,66	5,63
	0,025	12,22	10,65	9,98	9,60	9,36	9,20	9,07	8,98	8,90	8,84	8,66	8,56	8,50	8,46	8,41	8,36	8,32	8,26
	0,01	21,20	18,00	16,69	15,98	15,52	15,21	14,98	14,80	14,66	14,55	14,20	14,02	13,91	13,84	13,75	13,65	13,58	13,47
5	0,10	4,06	3,78	3,62	3,52	3,45	3,40	3,37	3,34	3,32	3,30	3,324	3,21	3,19	3,17	3,16	3,14	3,13	3,11
	0,05	6,61,	5,79	5,41	5,19	5,05	4,95	4,88	4,82	4,77	4,74	4,62	4,56	4,52	4,50	4,46	4,43	4,41	4,37
	0,025	10,01	8,43	7,76	7,39	7,15	6,98	6,85	6,76	6,68	6,62	6,43	6,33	6,27	6,23	6,18	6,12	6,08	6,02
	0,01	16,26	13,27	12,06	11,39	10,97	10,67	10,46	10,29	10,16	10,05	9,72	9,55	9,45	9,38	9,29	9,20	9,13	9,03

Tabela 4 Distribuição *F* (continuação)

Graus de Liberdade do | Área na Cauda | Graus de Liberdade do Numerador

Graus de Liberdade do Denominador	Área na Cauda Superior	Graus de Liberdade do Numerador																	
		1	2	3	4	5	6	7	8	9	10	15	20	25	30	40	60	100	1.000
6	0,10	3,78	3,46	3,29	3,18	3,11	3,05	3,01	2,98	2,96	2,94	2,87	2,84	2,81	2,80	2,78	2,76	2,75	2,72
	0,05	5,99	5,14	4,76	4,53	4,39	4,28	4,21	4,15	4,10	4,06	3,94	3,87	3,83	3,81	3,77	3,74	3,71	3,67
	0,025	8,81	7,26	6,60	6,23	5,99	5,82	5,70	5,60	5,52	5,46	5,27	5,17	5,11	5,07	5,01	4,96	4,92	4,86
	0,01	13,75	10,92	9,78	9,15	8,75	8,47	8,26	8,10	7,98	7,87	7,56	7,40	7,30	7,23	7,14	7,06	6,99	6,89
7	0,10	3,59	3,26	3,07	2,96	2,88	2,83	2,78	2,75	2,72	2,70	2,63	2,59	2,57	2,56	2,54	2,51	2,50	2,47
	0,05	5,59	4,74	4,35	4,12	3,97	3,87	3,79	3,73	3,68	3,64	3,51	3,44	3,40	3,38	3,34	3,30	3,27	3,23
	0,025	8,07	6,54	5,89	5,52	5,29	5,12	4,99	4,90	4,82	4,76	4,57	4,47	4,40	4,36	4,31	4,25	4,21	4,15
	0,01	12,25	9,55	8,45	7,85	7,46	7,19	6,99	6,84	6,72	6,62	6,31	6,16	6,06	5,99	5,91	5,82	5,75	5,66
8	0,10	3,46	3,11	2,92	2,81	2,73	2,67	2,62	2,59	2,56	2,54	2,46	2,42	2,40	2,38	2,36	2,34	2,32	2,30
	0,05	5,32	4,46	4,07	3,84	3,69	3,58	3,50	3,44	3,39	3,35	3,22	3,15	3,11	3,08	3,04	3,01	2,97	2,93
	0,025	7,57	6,06	5,42	5,05	4,82	4,65	4,53	4,43	4,36	4,30	4,10	4,00	3,94	3,89	3,84	3,78	3,74	3,68
	0,01	11,26	8,65	7,59	7,01	6,63	6,37	6,18	6,03	5,91	5,81	5,52	5,36	5,26	5,20	5,12	5,03	4,96	4,87
9	0,10	3,36	3,01	2,81	2,69	2,61	2,55	2,51	2,47	2,44	2,42	2,34	2,30	2,27	2,25	2,23	2,21	2,19	2,16
	0,05	5,12	4,26	3,86	3,63	3,48	3,37	3,29	3,23	3,18	3,14	3,01	2,94	2,89	2,86	2,83	2,79	2,76	2,71
	0,025	7,21	5,71	5,08	4,72	4,48	4,32	4,20	4,10	4,03	3,96	3,77	3,67	3,60	3,56	3,51	3,45	3,40	3,34
	0,01	10,56	8,02	6,99	6,42	6,06	5,80	5,61	5,47	5,35	5,26	4,96	4,81	4,71	4,65	4,57	4,48	4,41	4,32
10	0,10	3,29	2,92	2,73	2,61	2,52	2,46	2,41	2,38	2,35	2,32	2,24	2,20	2,17	2,16	2,13	2,11	2,09	2,06
	0,05	4,96	4,10	3,71	3,48	3,33	3,22	3,14	3,07	3,02	2,98	2,85	2,77	2,73	2,70	2,66	2,62	2,59	2,54
	0,025	6,94	5,46	4,83	4,47	4,24	4,07	3,95	3,85	3,78	3,72	3,52	3,42	3,35	3,31	3,26	3,20	3,15	3,09
	0,01	10,04	7,56	6,55	5,99	5,64	5,39	5,20	5,06	4,94	4,85	4,56	4,41	4,31	4,25	4,17	4,08	4,01	3,92
11	0,10	3,23	2,86	2,66	2,54	2,45	2,39	2,34	2,30	2,27	2,25	2,17	2,12	2,10	2,08	2,05	2,03	2,01	1,98
	0,05	4,84	3,98	3,59	3,36	3,20	3,09	3,01	2,95	2,90	2,85	2,72	2,65	2,60	2,57	2,53	2,49	2,46	2,41
	0,025	6,72	5,26	4,63	4,28	4,04	3,88	3,76	3,66	3,59	3,53	3,33	3,23	3,16	3,12	3,06	3,00	2,96	2,89
	0,01	9,65	7,21	6,22	5,67	5,32	5,07	4,89	4,74	4,63	4,54	4,25	4,10	4,01	3,94	3,86	3,78	3,71	3,61
12	0,10	3,18	2,81	2,61	2,48	2,39	2,33	2,28	2,24	2,21	2,19	2,10	2,06	2,03	2,01	1,99	1,96	1,94	1,91
	0,05	4,75	3,89	3,49	3,26	3,11	3,00	2,91	2,85	2,80	2,75	2,62	2,54	2,50	2,47	2,43	2,38	2,35	2,30
	0,025	6,55	5,10	4,47	4,12	3,89	3,73	3,61	3,51	3,44	3,37	3,18	3,07	3,01	2,96	2,91	2,85	2,80	2,73
	0,01	9,33	6,93	5,95	5,41	5,06	4,82	4,64	4,50	4,39	4,30	4,01	3,86	3,76	3,70	3,62	3,54	3,47	3,37
13	0,10	3,14	2,76	2,56	2,43	2,35	2,28	2,23	2,20	2,16	2,14	2,05	2,01	1,98	1,96	1,93	1,90	1,88	1,85
	0,05	4,67	3,81	3,41	3,18	3,03	2,92	2,83	2,77	2,71	2,67	2,53	2,46	2,41	2,38	2,34	2,30	2,26	2,21
	0,025	6,41	4,97	4,35	4,00	3,77	3,60	3,48	3,39	3,31	3,25	3,05	2,95	2,88	2,84	2,78	2,72	2,67	2,60
	0,01	9,07	6,70	5,74	5,21	4,86	4,62	4,44	4,30	4,19	4,10	3,82	3,66	3,57	3,51	3,43	3,34	3,27	3,18
14	0,10	3,10	2,73	2,52	2,39	2,31	2,24	2,19	2,15	2,12	2,10	2,01	1,96	1,93	1,99	1,89	1,86	1,83	1,80
	0,05	4,60	3,74	3,34	3,11	2,96	2,85	2,76	2,70	2,65	2,60	2,46	2,39	2,34	2,31	2,27	2,22	2,19	2,14
	0,025	6,30	4,86	4,24	3,89	3,66	3,50	3,38	3,29	3,21	3,15	2,95	2,84	2,78	2,73	2,67	2,61	2,56	2,50
	0,01	8,86	6,51	5,56	5,04	4,69	4,46	4,28	4,14	4,03	3,94	3,66	3,51	3,41	3,35	3,27	3,18	3,11	3,02
15	0,10	3,07	2,70	2,49	2,36	2,27	2,21	2,16	2,12	2,09	2,06	1,97	1,92	1,89	1,87	1,85	1,82	1,79	1,76

Tabela 4 Distribuição *F* (continuação)

Graus de Liberdade do Denominador	Área na Cauda Superior	Graus de Liberdade do Numerador																	
		1	2	3	4	5	6	7	8	9	10	15	20	25	30	40	60	100	1.000
	0,05	4,54	3,68	3,29	3,06	2,90	2,79	2,71	2,64	2,59	2,54	2,40	2,33	2,28	2,25	2,20	2,16	2,12	2,07
	0,025	6,20	4,77	4,15	3,80	3,58	3,41	3,29	3,20	3,12	3,06	2,86	2,76	2,69	2,64	2,59	2,52	2,47	2,40
	0,01	8,68	6,36	5,42	4,89	4,56	4,32	4,14	4,00	3,89	3,80	3,52	3,37	3,28	3,21	3,13	3,05	2,98	2,88
16	0,10	3,05	2,67	2,46	2,33	2,24	2,18	2,13	2,09	2,06	2,03	1,94	1,89	1,86	1,84	1,81	1,78	1,76	1,72
	0,05	4,49	3,63	3,24	3,01	2,85	2,74	2,66	2,59	2,54	2,49	2,35	2,28	2,23	2,19	2,15	2,11	2,07	2,02
	0,025	6,12	4,69	4,08	3,73	3,50	3,34	3,22	3,12	3,05	2,99	2,79	2,68	2,61	2,57	2,51	2,45	2,40	2,32
	0,01	8,53	6,23	5,29	4,77	4,44	4,20	4,03	3,89	3,78	3,69	3,41	3,26	3,16	3,10	3,02	2,93	2,86	2,76
17	0,10	3,03	2,64	2,44	2,31	2,22	2,15	2,10	2,06	2,03	2,00	1,91	1,86	1,83	1,81	1,78	1,75	1,73	1,69
	0,05	4,45	3,59	3,20	2,96	2,81	2,70	2,61	2,55	2,49	2,45	2,31	2,23	2,18	2,15	2,10	2,06	2,02	1,97
	0,025	6,04	4,62	4,01	3,66	3,44	3,28	3,16	3,06	2,98	2,92	2,72	2,62	2,55	2,50	2,44	2,38	2,33	2,26
	0,01	8,40	6,11	5,19	4,67	4,34	4,10	3,93	3,79	3,68	3,59	3,31	3,16	3,07	3,00	2,92	2,83	2,76	2,66
18	0,10	3,01	2,62	2,42	2,29	2,20	2,13	2,08	2,04	2,00	1,98	1,89	1,84	1,80	1,78	1,75	1,72	1,70	1,66
	0,05	4,41	3,55	3,16	2,93	2,77	2,66	2,58	2,51	2,46	2,41	2,27	2,19	2,14	2,11	2,06	2,02	1,98	1,92
	0,025	5,98	4,56	3,95	3,61	3,38	3,22	3,10	3,01	2,93	2,87	2,67	2,56	2,49	2,44	2,38	2,32	2,27	2,20
	0,01	8,29	6,01	5,09	4,58	4,25	4,01	3,84	3,71	3,60	3,51	3,23	3,08	2,98	2,92	2,84	2,75	2,68	2,58
19	0,10	2,99	2,61	2,40	2,27	2,18	2,11	2,06	2,02	1,98	1,96	1,86	1,81	1,78	1,76	1,73	1,70	1,67	1,64
	0,05	4,38	3,52	3,13	2,90	2,74	2,63	2,54	2,48	2,42	2,38	2,23	2,16	2,11	2,07	2,03	1,98	1,94	1,88
	0,025	5,92	4,51	3,90	3,56	3,33	3,17	3,05	2,96	2,88	2,82	2,62	2,51	2,44	2,39	2,33	2,27	2,22	2,14
	0,01	8,18	5,93	5,01	4,50	4,17	3,94	3,77	3,63	3,52	3,43	3,15	3,00	2,91	2,84	2,76	2,67	2,60	2,50
20	0,10	2,97	2,59	2,38	2,25	2,16	2,09	2,04	2,00	1,96	1,94	1,84	1,79	1,76	1,74	1,71	1,68	1,65	1,61
	0,05	4,35	3,49	3,10	2,87	2,71	2,60	2,51	2,45	2,39	2,35	2,20	2,12	2,07	2,04	1,99	1,95	1,91	1,85
	0,025	5,87	4,46	3,86	3,51	3,29	3,13	3,01	2,91	2,84	2,77	2,57	2,46	2,40	2,35	2,29	2,22	2,17	2,09
	0,01	8,10	5,85	4,94	4,43	4,10	3,87	3,70	3,56	3,46	3,37	3,09	2,94	2,84	2,78	2,69	2,61	2,54	2,43
21	0,10	2,96	2,57	2,36	2,23	2,14	2,08	2,02	1,98	1,95	1,92	1,83	1,78	1,74	1,72	1,69	1,66	1,63	1,59
	0,05	4,32	3,47	3,07	2,84	2,68	2,57	2,49	2,42	2,37	2,32	2,18	2,10	2,05	2,01	1,96	1,92	1,88	1,82
	0,025	5,83	4,42	3,82	3,48	3,25	3,09	2,97	2,87	2,80	2,73	2,53	2,42	2,36	2,31	2,25	2,18	2,13	2,05
	0,01	8,02	5,78	4,87	4,37	4,04	3,81	3,64	3,51	3,40	3,31	3,03	2,88	2,79	2,72	2,64	2,55	2,48	2,37
22	0,10	2,95	2,56	2,35	2,22	2,13	2,06	2,01	1,97	1,93	1,90	1,81	1,76	1,73	1,70	1,67	1,64	1,61	1,57
	0,05	4,30	3,44	3,05	2,82	2,66	2,55	2,46	2,40	2,34	2,30	2,15	2,07	2,02	1,98	1,94	1,89	1,85	1,79
	0,025	5,79	4,38	3,78	3,44	3,22	3,05	2,93	2,84	2,76	2,70	2,50	2,39	2,32	2,27	2,21	2,14	2,09	2,01
	0,01	7,95	5,72	4,82	4,31	3,99	3,76	3,59	3,45	3,35	3,26	2,98	2,83	2,73	2,67	2,58	2,50	2,42	2,32
23	0,10	2,94	2,55	2,34	2,21	2,11	2,05	1,99	1,95	1,92	1,89	1,80	1,74	1,71	1,69	1,66	1,62	1,59	1,55
	0,05	4,28	3,42	3,03	2,80	2,64	2,53	2,44	2,37	2,32	2,27	2,13	2,05	2,00	1,96	1,91	1,86	1,82	1,76
	0,025	5,75	4,35	3,75	3,41	3,18	3,02	2,90	2,81	2,73	2,67	2,47	2,36	2,29	2,24	2,18	2,11	2,06	1,98
	0,01	7,88	5,66	4,76	4,26	3,94	3,71	3,54	3,41	3,30	3,21	2,93	2,78	2,69	2,62	2,54	2,45	2,37	2,27
24	0,10	2,93	2,54	2,33	2,19	2,10	2,04	1,98	1,94	1,91	1,88	1,78	1,73	1,70	1,67	1,64	1,61	1,58	1,54
	0,05	4,26	3,40	3,01	2,78	2,62	2,51	2,42	2,36	2,30	2,25	2,11	2,03	1,97	1,94	1,89	1,84	1,80	1,74
	0,025	5,72	4,32	3,72	3,38	3,15	2,99	2,87	2,78	2,70	2,64	2,44	2,33	2,26	2,21	2,15	2,08	2,02	1,94
	0,01	7,82	5,61	4,72	4,22	3,90	3,67	3,50	3,36	3,26	3,17	2,89	2,74	2,64	2,58	2,49	2,40	2,33	2,22

Tabela 4 Distribuição F (continuação)

Graus de Liberdade do Denominador	Área na Cauda Superior	Graus de Liberdade do Numerador																	
		1	2	3	4	5	6	7	8	9	10	15	20	25	30	40	60	100	1.000
25	0,10	2,92	2,53	2,32	2,18	2,09	2,02	1,97	1,93	1,89	1,87	1,77	1,72	1,68	1,66	1,63	1,59	1,56	1,52
	0,05	4,24	3,39	2,99	2,76	2,60	2,49	2,40	2,34	2,28	2,24	2,09	2,01	1,96	1,92	1,87	1,82	1,78	1,72
	0,025	5,69	4,29	3,69	3,35	3,13	2,97	2,85	2,75	2,68	2,61	2,41	2,30	2,23	2,18	2,12	2,05	2,00	1,91
	0,01	7,77	5,57	4,68	4,18	3,85	3,63	3,46	3,32	3,22	3,13	2,85	2,70	2,60	2,54	2,45	2,36	2,29	2,18
26	0,10	2,91	2,52	2,31	2,17	2,08	2,01	1,96	1,92	1,88	1,86	1,76	1,71	1,67	1,65	1,61	1,58	1,55	1,51
	0,05	4,23	3,37	2,98	2,74	2,59	2,47	2,39	2,32	2,27	2,22	2,07	1,99	1,94	1,90	1,85	1,80	1,76	1,70
	0,025	5,66	4,27	3,67	3,33	3,10	2,94	2,82	2,73	2,65	2,59	2,39	2,28	2,21	2,16	2,09	2,03	1,97	1,89
	0,01	7,72	5,53	4,64	4,14	3,82	3,59	3,42	3,29	3,18	3,09	2,81	2,66	2,57	2,50	2,42	2,33	2,25	2,14
27	0,10	2,90	2,51	2,30	2,17	2,07	2,00	1,95	1,91	1,87	1,85	1,75	1,70	1,66	1,64	1,60	1,57	1,54	1,50
	0,05	4,21	3,35	2,96	2,73	2,57	2,46	2,37	2,31	2,25	2,20	2,06	1,97	1,92	1,88	1,84	1,79	1,74	1,68
	0,025	5,63	4,24	3,65	3,31	3,08	2,92	2,80	2,71	2,63	2,57	2,36	2,25	2,18	2,13	2,07	2,00	1,94	1,86
	0,01	7,68	5,49	4,60	4,11	3,78	3,56	3,39	3,26	3,15	3,06	2,78	2,63	2,54	2,47	2,38	2,29	2,22	2,11
28	0,10	2,89	2,50	2,29	2,16	2,06	2,00	1,94	1,90	1,87	1,84	1,74	1,69	1,65	1,63	1,59	1,56	1,53	1,48
	0,05	4,20	3,34	2,95	2,71	2,56	2,45	2,36	2,29	2,24	2,19	2,04	1,96	1,91	1,87	1,82	1,77	1,73	1,66
	0,025	5,61	4,22	3,63	3,29	3,06	2,90	2,78	2,69	2,61	2,55	2,34	2,23	2,16	2,11	2,05	1,98	1,92	1,84
	0,01	7,64	5,45	4,57	4,07	3,75	3,53	3,36	3,23	3,12	3,03	2,75	2,60	2,51	2,44	2,35	2,26	2,19	2,08
29	0,10	2,89	2,50	2,28	2,15	2,06	1,99	1,93	1,89	1,86	1,83	1,73	1,68	1,64	1,62	1,58	1,55	1,52	1,47
	0,05	4,18	3,33	2,93	2,70	2,55	2,43	2,35	2,28	2,22	2,18	2,03	1,94	1,89	1,85	1,81	1,75	1,71	1,65
	0,025	5,59	4,20	3,61	3,27	3,04	2,88	2,76	2,67	2,59	2,53	2,32	2,21	2,14	2,09	2,03	1,96	1,90	1,82
	0,01	7,60	5,42	4,54	4,04	3,73	3,50	3,33	3,20	3,09	3,00	2,73	2,57	2,48	2,41	2,33	2,23	2,16	2,05
30	0,10	2,88	2,49	2,28	2,14	2,05	1,98	1,93	1,88	1,85	1,82	1,72	1,67	1,63	1,61	1,57	1,54	1,51	1,46
	0,05	4,17	3,32	2,92	2,69	2,53	2,42	2,33	2,27	2,21	2,16	2,01	1,93	1,88	1,84	1,79	1,74	1,70	1,63
	0,025	5,57	4,18	3,59	3,25	3,03	2,87	2,75	2,65	2,57	2,51	2,31	2,20	2,12	2,07	2,01	1,94	1,88	1,80
	0,01	7,56	5,39	4,51	4,02	3,70	3,47	3,30	3,17	3,07	2,98	2,70	2,55	2,45	2,39	2,30	2,21	2,13	2,02
40	0,10	2,84	2,44	2,23	2,09	2,00	1,93	1,87	1,83	1,79	1,76	1,66	1,61	1,57	1,54	1,51	1,47	1,43	1,38
	0,05	4,08	3,23	2,84	2,61	2,45	2,34	2,25	2,18	2,12	2,08	1,92	1,84	1,78	1,74	1,69	1,64	1,59	1,52
	0,025	5,42	4,05	3,46	3,13	2,90	2,74	2,62	2,53	2,45	2,39	2,18	2,07	1,99	1,94	1,88	1,80	1,74	1,65
	0,01	7,31	5,18	4,31	3,83	3,51	3,29	3,12	2,99	2,89	2,80	2,52	2,37	2,27	2,20	2,11	2,02	1,94	1,82
60	0,10	2,79	2,39	2,18	2,04	1,95	1,87	1,82	1,77	1,74	1,71	1,60	1,54	1,50	1,48	1,44	1,40	1,36	1,30
	0,05	4,00	3,15	2,76	2,53	2,37	2,25	2,17	2,10	2,04	1,99	1,84	1,75	1,69	1,65	1,59	1,53	1,48	1,40
	0,025	5,29	3,93	3,34	3,01	2,79	2,63	2,51	2,41	2,33	2,27	2,06	1,94	1,87	1,82	1,74	1,67	1,60	1,49
	0,01	7,08	4,98	4,13	3,65	3,34	3,12	2,95	2,82	2,72	2,63	2,35	2,20	2,10	2,03	1,94	1,84	1,75	1,62
100	0,10	2,76	2,36	2,14	2,00	1,91	1,83	1,78	1,73	1,69	1,66	1,56	1,49	1,45	1,42	1,38	1,34	1,29	1,22
	0,05	3,94	3,09	2,70	2,46	2,31	2,19	2,10	2,03	1,97	1,93	1,77	1,68	1,62	1,57	1,52	1,45	1,39	1,30
	0,025	5,18	3,83	3,25	2,92	2,70	2,54	2,42	2,32	2,24	2,18	1,97	1,85	1,77	1,71	1,64	1,56	1,48	1,36
	0,01	6,90	4,82	3,98	3,51	3,21	2,99	2,82	2,69	2,59	2,50	2,22	2,07	1,97	1,89	1,80	1,69	1,60	1,45
1.000	0,10	2,71	2,31	2,09	1,95	1,85	1,78	1,72	1,68	1,64	1,61	1,49	1,43	1,38	1,35	1,30	1,25	1,20	1,08
	0,05	3,85	3,00	2,61	2,38	2,22	2,11	2,02	1,95	1,89	1,84	1,68	1,58	1,52	1,47	1,41	1,33	1,26	1,11
	0,025	5,04	3,70	3,13	2,80	2,58	2,42	2,30	2,20	2,13	2,06	1,85	1,72	1,64	1,58	1,50	1,41	1,32	1,13
	0,01	6,66	4,63	3,80	3,34	3,04	2,82	2,66	2,53	2,43	2,34	2,06	1,90	1,79	1,72	1,61	1,50	1,38	1,16

Tabela 5 Probabilidades Binomiais

Os registros na tabela fornecem a probabilidade de x sucessos em n ensaios de um experimento binomial, em que p é a probabilidade de sucesso em um ensaio. Por exemplo, com seis ensaios e $p = 0{,}05$, a probabilidade de dois sucessos é 0,0305.

		p								
n	x	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
2	0	0,9801	0,9604	0,9409	0,9216	0,9025	0,8836	0,8649	0,8464	0,8281
	1	0,0198	0,0392	0,0582	0,0768	0,0950	0,1128	0,1302	0,1472	0,1638
	2	0,0001	0,0004	0,0009	0,0016	0,0025	0,0036	0,0049	0,0064	0,0081
3	0	0,9703	0,9412	0,9127	0,8847	0,8574	0,8306	0,8044	0,7787	0,7536
	1	0,0294	0,0576	0,0847	0,1106	0,1354	0,1590	0,1816	0,2031	0,2236
	2	0,0003	0,0012	0,0026	0,0046	0,0071	0,0102	0,0137	0,0177	0,0221
	3	0,0000	0,0000	0,0000	0,0001	0,0001	0,0002	0,0003	0,0005	0,0007
4	0	0,9606	0,9224	0,8853	0,8493	0,8145	0,7807	0,7481	0,7164	0,6857
	1	0,0388	0,0753	0,1095	0,1416	0,1715	0,1993	0,2252	0,2492	0,2713
	2	0,0006	0,0023	0,0051	0,0088	0,0135	0,0191	0,0254	0,0325	0,0402
	3	0,0000	0,0000	0,0001	0,0002	0,0005	0,0008	0,0013	0,0019	0,0027
	4	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001
5	0	0,9510	0,9039	0,8587	0,8154	0,7738	0,7339	0,6957	0,6591	0,6240
	1	0,0480	0,0922	0,1328	0,1699	0,2036	0,2342	0,2618	0,2866	0,3086
	2	0,0010	0,0038	0,0082	0,0142	0,0214	0,0299	0,0394	0,0498	0,0610
	3	0,0000	0,0001	0,0003	0,0006	0,0011	0,0019	0,0030	0,0043	0,0060
	4	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001	0,0001	0,0002	0,0003
	5	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
6	0	0,9415	0,8858	0,8330	0,7828	0,7351	0,6899	0,6470	0,6064	0,5679
	1	0,0571	0,1085	0,1546	0,1957	0,2321	0,2642	0,2922	0,3164	0,3370
	2	0,0014	0,0055	0,0120	0,0204	0,0305	0,0422	0,0550	0,0688	0,0833
	3	0,0000	0,0002	0,0005	0,0011	0,0021	0,0036	0,0055	0,0080	0,0110
	4	0,0000	0,0000	0,0000	0,0000	0,0001	0,0002	0,0003	0,0005	0,0008
	5	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
	6	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
7	0	0,9321	0,8681	0,8080	0,7514	0,6983	0,6485	0,6017	0,5578	0,5168
	1	0,0659	0,1240	0,1749	0,2192	0,2573	0,2897	0,3170	0,3396	0,3578
	2	0,0020	0,0076	0,0162	0,0274	0,0406	0,0555	0,0716	0,0886	0,1061
	3	0,0000	0,0003	0,0008	0,0019	0,0036	0,0059	0,0090	0,0128	0,0175
	4	0,0000	0,0000	0,0000	0,0001	0,0002	0,0004	0,0007	0,0011	0,0017
	5	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001	0,0001
	6	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
	7	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
8	0	0,9227	0,8508	0,7837	0,7214	0,6634	0,6096	0,5596	0,5132	0,4703
	1	0,0746	0,1389	0,1939	0,2405	0,2793	0,3113	0,3370	0,3570	0,3721
	2	0,0026	0,0099	0,0210	0,0351	0,0515	0,0695	0,0888	0,1087	0,1288
	3	0,0001	0,0004	0,0013	0,0029	0,0054	0,0089	0,0134	0,0189	0,0255
	4	0,0000	0,0000	0,0001	0,0002	0,0004	0,0007	0,0013	0,0021	0,0031
	5	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001	0,0001	0,0002
	6	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
	7	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
	8	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000

Tabela 5 Probabilidades Binomiais (continuação)

n	x	p								
		0,01	**0,02**	**0,03**	**0,04**	**0,05**	**0,06**	**0,07**	**0,08**	**0,09**
9	0	0,9135	0,8337	0,7602	0,6925	0,6302	0,5730	0,5204	0,4722	0,4279
	1	0,0830	0,1531	0,2116	0,2597	0,2985	0,3292	0,3525	0,3695	0,3809
	2	0,0034	0,0125	0,0262	0,0433	0,0629	0,0840	0,1061	0,1285	0,1507
	3	0,0001	0,0006	0,0019	0,0042	0,0077	0,0125	0,0186	0,0261	0,0348
	4	0,0000	0,0000	0,0001	0,0003	0,0006	0,0012	0,0021	0,0034	0,0052
	5	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001	0,0002	0,0003	0,0005
	6	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
	7	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
	8	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
	9	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
10	0	0,9044	0,8171	0,7374	0,6648	0,5987	0,5386	0,4840	0,4344	0,3894
	1	0,0914	0,1667	0,2281	0,2770	0,3151	0,3438	0,3643	0,3777	0,3851
	2	0,0042	0,0153	0,0317	0,0519	0,0746	0,0988	0,1234	0,1478	0,1714
	3	0,0001	0,0008	0,0026	0,0058	0,0105	0,0168	0,0248	0,0343	0,0452
	4	0,0000	0,0000	0,0001	0,0004	0,0010	0,0019	0,0033	0,0052	0,0078
	5	0,0000	0,0000	0,0000	0,0000	0,0001	0,0001	0,0003	0,0005	0,0009
	6	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001
	7	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
	8	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
	9	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
	10	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
12	0	0,8864	0,7847	0,6938	0,6127	0,5404	0,4759	0,4186	0,3677	0,3225
	1	0,1074	0,1922	0,2575	0,3064	0,3413	0,3645	0,3781	0,3837	0,3827
	2	0,0060	0,0216	0,0438	0,0702	0,0988	0,1280	0,1565	0,1835	0,2082
	3	0,0002	0,0015	0,0045	0,0098	0,0173	0,0272	0,0393	0,0532	0,0686
	4	0,0000	0,0001	0,0003	0,0009	0,0021	0,0039	0,0067	0,0104	0,0153
	5	0,0000	0,0000	0,0000	0,0001	0,0002	0,0004	0,0008	0,0014	0,0024
	6	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001	0,0001	0,0003
	7	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
	8	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
	9	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
	10	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
	11	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
	12	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
15	0	0,8601	0,7386	0,6333	0,5421	0,4633	0,3953	0,3367	0,2863	0,2430
	1	0,1303	0,2261	0,2938	0,3388	0,3658	0,3785	0,3801	0,3734	0,3605
	2	0,0092	0,0323	0,0636	0,0988	0,1348	0,1691	0,2003	0,2273	0,2496
	3	0,0004	0,0029	0,0085	0,0178	0,0307	0,0468	0,0653	0,0857	0,1070
	4	0,0000	0,0002	0,0008	0,0022	0,0049	0,0090	0,0148	0,0223	0,0317
	5	0,0000	0,0000	0,0001	0,0002	0,0006	0,0013	0,0024	0,0043	0,0069
	6	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001	0,0003	0,0006	0,0011
	7	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001	0,0001
	8	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
	9	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
	10	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
	11	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
	12	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
	13	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
	14	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
	15	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000

Tabela 5 Probabilidades Binomiais (continuação)

		p								
n	*x*	**0,01**	**0,02**	**0,03**	**0,04**	**0,05**	**0,06**	**0,07**	**0,08**	**0,09**
18	0	0,8345	0,6951	0,5780	0,4796	0,3972	0,3283	0,2708	0,2229	0,1831
	1	0,1517	0,2554	0,3217	0,3597	0,3763	0,3772	0,3669	0,3489	0,3260
	2	0,0130	0,0443	0,0846	0,1274	0,1683	0,2047	0,2348	0,2579	0,2741
	3	0,0007	0,0048	0,0140	0,0283	0,0473	0,0697	0,0942	0,1196	0,1446
	4	0,0000	0,0004	0,0016	0,0044	0,0093	0,0167	0,0266	0,0390	0,0536
	5	0,0000	0,0000	0,0001	0,0005	0,0014	0,0030	0,0056	0,0095	0,0148
	6	0,0000	0,0000	0,0000	0,0000	0,0002	0,0004	0,0009	0,0018	0,0032
	7	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001	0,0003	0,0005
	8	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001
	9	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
	10	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
	11	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
	12	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
	13	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
	14	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
	15	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
	16	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
	17	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
	18	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
20	0	0,8179	0,6676	0,5438	0,4420	0,3585	0,2901	0,2342	0,1887	0,1516
	1	0,1652	0,2725	0,3364	0,3683	0,3774	0,3703	0,3526	0,3282	0,3000
	2	0,0159	0,0528	0,0988	0,1458	0,1887	0,2246	0,2521	0,2711	0,2818
	3	0,0010	0,0065	0,0183	0,0364	0,0596	0,0860	0,1139	0,1414	0,1672
	4	0,0000	0,0006	0,0024	0,0065	0,0133	0,0233	0,0364	0,0523	0,0703
	5	0,0000	0,0000	0,0002	0,0009	0,0022	0,0048	0,0088	0,0145	0,0222
	6	0,0000	0,0000	0,0000	0,0001	0,0003	0,0008	0,0017	0,0032	0,0055
	7	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001	0,0002	0,0005	0,0011
	8	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001	0,0002
	9	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
	10	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
	11	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
	12	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
	13	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
	14	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
	15	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
	16	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
	17	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
	18	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
	19	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
	20	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000

Tabela 5 Probabilidades Binomiais (continuação)

		p								
n	*x*	0,10	0,15	0,20	0,25	0,30	0,35	0,40	0,45	0,50
2	0	0,8100	0,7225	0,6400	0,5625	0,4900	0,4225	0,3600	0,3025	0,2500
	1	0,1800	0,2550	0,3200	0,3750	0,4200	0,4550	0,4800	0,4950	0,5000
	2	0,0100	0,0225	0,0400	0,0625	0,0900	0,1225	0,1600	0,2025	0,2500
3	0	0,7290	0,6141	0,5120	0,4219	0,3430	0,2746	0,2160	0,1664	0,1250
	1	0,2430	0,3251	0,3840	0,4219	0,4410	0,4436	0,4320	0,4084	0,3750
	2	0,0270	0,0574	0,0960	0,1406	0,1890	0,2389	0,2880	0,3341	0,3750
	3	0,0010	0,0034	0,0080	0,0156	0,0270	0,0429	0,0640	0,0911	0,1250
4	0	0,6561	0,5220	0,4096	0,3164	0,2401	0,1785	0,1296	0,0915	0,0625
	1	0,2916	0,3685	0,4096	0,4219	0,4116	0,3845	0,3456	0,2995	0,2500
	2	0,0486	0,0975	0,1536	0,2109	0,2646	0,3105	0,3456	0,3675	0,3750
	3	0,0036	0,0115	0,0256	0,0469	0,0756	0,1115	0,1536	0,2005	0,2500
	4	0,0001	0,0005	0,0016	0,0039	0,0081	0,0150	0,0256	0,0410	0,0625
5	0	0,5905	0,4437	0,3277	0,2373	0,1681	0,1160	0,0778	0,0503	0,0312
	1	0,3280	0,3915	0,4096	0,3955	0,3602	0,3124	0,2592	0,2059	0,1562
	2	0,0729	0,1382	0,2048	0,2637	0,3087	0,3364	0,3456	0,3369	0,3125
	3	0,0081	0,0244	0,0512	0,0879	0,1323	0,1811	0,2304	0,2757	0,3125
	4	0,0004	0,0022	0,0064	0,0146	0,0284	0,0488	0,0768	0,1128	0,1562
	5	0,0000	0,0001	0,0003	0,0010	0,0024	0,0053	0,0102	0,0185	0,0312
6	0	0,5314	0,3771	0,2621	0,1780	0,1176	0,0754	0,0467	0,0277	0,0156
	1	0,3543	0,3993	0,3932	0,3560	0,3025	0,2437	0,1866	0,1359	0,0938
	2	0,0984	0,1762	0,2458	0,2966	0,3241	0,3280	0,3110	0,2780	0,2344
	3	0,0146	0,0415	0,0819	0,1318	0,1852	0,2355	0,2765	0,3032	0,3125
	4	0,0012	0,0055	0,0154	0,0330	0,0595	0,0951	0,1382	0,1861	0,2344
	5	0,0001	0,0004	0,0015	0,0044	0,0102	0,0205	0,0369	0,0609	0,0938
	6	0,0000	0,0000	0,0001	0,0002	0,0007	0,0018	0,0041	0,0083	0,0156
7	0	0,4783	0,3206	0,2097	0,1335	0,0824	0,0490	0,0280	0,0152	0,0078
	1	0,3720	0,3960	0,3670	0,3115	0,2471	0,1848	0,1306	0,0872	0,0547
	2	0,1240	0,2097	0,2753	0,3115	0,3177	0,2985	0,2613	0,2140	0,1641
	3	0,0230	0,0617	0,1147	0,1730	0,2269	0,2679	0,2903	0,2918	0,2734
	4	0,0026	0,0109	0,0287	0,0577	0,0972	0,1442	0,1935	0,2388	0,2734
	5	0,0002	0,0012	0,0043	0,0115	0,0250	0,0466	0,0774	0,1172	0,1641
	6	0,0000	0,0001	0,0004	0,0013	0,0036	0,0084	0,0172	0,0320	0,0547
	7	0,0000	0,0000	0,0000	0,0001	0,0002	0,0006	0,0016	0,0037	0,0078
8	0	0,4305	0,2725	0,1678	0,1001	0,0576	0,0319	0,0168	0,0084	0,0039
	1	0,3826	0,3847	0,3355	0,2670	0,1977	0,1373	0,0896	0,0548	0,0312
	2	0,1488	0,2376	0,2936	0,3115	0,2965	0,2587	0,2090	0,1569	0,1094
	3	0,0331	0,0839	0,1468	0,2076	0,2541	0,2786	0,2787	0,2568	0,2188
	4	0,0046	0,0185	0,0459	0,0865	0,1361	0,1875	0,2322	0,2627	0,2734
	5	0,0004	0,0026	0,0092	0,0231	0,0467	0,0808	0,1239	0,1719	0,2188
	6	0,0000	0,0002	0,0011	0,0038	0,0100	0,0217	0,0413	0,0703	0,1094
	7	0,0000	0,0000	0,0001	0,0004	0,0012	0,0033	0,0079	0,0164	0,0313
	8	0,0000	0,0000	0,0000	0,0000	0,0001	0,0002	0,0007	0,0017	0,0039

Tabela 5 Probabilidades Binomiais (continuação)

		p								
n	x	0,10	0,15	0,20	0,25	0,30	0,35	0,40	0,45	0,50
9	0	0,3874	0,2316	0,1342	0,0751	0,0404	0,0207	0,0101	0,0046	0,0020
	1	0,3874	0,3679	0,3020	0,2253	0,1556	0,1004	0,0605	0,0339	0,0176
	2	0,1722	0,2597	0,3020	0,3003	0,2668	0,2162	0,1612	0,1110	0,0703
	3	0,0446	0,1069	0,1762	0,2336	0,2668	0,2716	0,2508	0,2119	0,1641
	4	0,0074	0,0283	0,0661	0,1168	0,1715	0,2194	0,2508	0,2600	0,2461
	5	0,0008	0,0050	0,0165	0,0389	0,0735	0,1181	0,1672	0,2128	0,2461
	6	0,0001	0,0006	0,0028	0,0087	0,0210	0,0424	0,0743	0,1160	0,1641
	7	0,0000	0,0000	0,0003	0,0012	0,0039	0,0098	0,0212	0,0407	0,0703
	8	0,0000	0,0000	0,0000	0,0001	0,0004	0,0013	0,0035	0,0083	0,0176
	9	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001	0,0003	0,0008	0,0020
10	0	0,3487	0,1969	0,1074	0,0563	0,0282	0,0135	0,0060	0,0025	0,0010
	1	0,3874	0,3474	0,2684	0,1877	0,1211	0,0725	0,0403	0,0207	0,0098
	2	0,1937	0,2759	0,3020	0,2816	0,2335	0,1757	0,1209	0,0763	0,0439
	3	0,0574	0,1298	0,2013	0,2503	0,2668	0,2522	0,2150	0,1665	0,1172
	4	0,0112	0,0401	0,0881	0,1460	0,2001	0,2377	0,2508	0,2384	0,2051
	5	0,0015	0,0085	0,0264	0,0584	0,1029	0,1536	0,2007	0,2340	0,2461
	6	0,0001	0,0012	0,0055	0,0162	0,0368	0,0689	0,1115	0,1596	0,2051
	7	0,0000	0,0001	0,0008	0,0031	0,0090	0,0212	0,0425	0,0746	0,1172
	8	0,0000	0,0000	0,0001	0,0004	0,0014	0,0043	0,0106	0,0229	0,0439
	9	0,0000	0,0000	0,0000	0,0000	0,0001	0,0005	0,0016	0,0042	0,0098
	10	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001	0,0003	0,0010
12	0	0,2824	0,1422	0,0687	0,0317	0,0138	0,0057	0,0022	0,0008	0,0002
	1	0,3766	0,3012	0,2062	0,1267	0,0712	0,0368	0,0174	0,0075	0,0029
	2	0,2301	0,2924	0,2835	0,2323	0,1678	0,1088	0,0639	0,0339	0,0161
	3	0,0853	0,1720	0,2362	0,2581	0,2397	0,1954	0,1419	0,0923	0,0537
	4	0,0213	0,0683	0,1329	0,1936	0,2311	0,2367	0,2128	0,1700	0,1208
	5	0,0038	0,0193	0,0532	0,1032	0,1585	0,2039	0,2270	0,2225	0,1934
	6	0,0005	0,0040	0,0155	0,0401	0,0792	0,1281	0,1766	0,2124	0,2256
	7	0,0000	0,0006	0,0033	0,0115	0,0291	0,0591	0,1009	0,1489	0,1934
	8	0,0000	0,0001	0,0005	0,0024	0,0078	0,0199	0,0420	0,0762	0,1208
	9	0,0000	0,0000	0,0001	0,0004	0,0015	0,0048	0,0125	0,0277	0,0537
	10	0,0000	0,0000	0,0000	0,0000	0,0002	0,0008	0,0025	0,0068	0,0161
	11	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001	0,0003	0,0010	0,0029
	12	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001	0,0002
15	0	0,2059	0,0874	0,0352	0,0134	0,0047	0,0016	0,0005	0,0001	0,0000
	1	0,3432	0,2312	0,1319	0,0668	0,0305	0,0126	0,0047	0,0016	0,0005
	2	0,2669	0,2856	0,2309	0,1559	0,0916	0,0476	0,0219	0,0090	0,0032
	3	0,1285	0,2184	0,2501	0,2252	0,1700	0,1110	0,0634	0,0318	0,0139
	4	0,0428	0,1156	0,1876	0,2252	0,2186	0,1792	0,1268	0,0780	0,0417
	5	0,0105	0,0449	0,1032	0,1651	0,2061	0,2123	0,1859	0,1404	0,0916
	6	0,0019	0,0132	0,0430	0,0917	0,1472	0,1906	0,2066	0,1914	0,1527
	7	0,0003	0,0030	0,0138	0,0393	0,0811	0,1319	0,1771	0,2013	0,1964
	8	0,0000	0,0005	0,0035	0,0131	0,0348	0,0710	0,1181	0,1647	0,1964
	9	0,0000	0,0001	0,0007	0,0034	0,0016	0,0298	0,0612	0,1048	0,1527
	10	0,0000	0,0000	0,0001	0,0007	0,0030	0,0096	0,0245	0,0515	0,0916
	11	0,0000	0,0000	0,0000	0,0001	0,0006	0,0024	0,0074	0,0191	0,0417
	12	0,0000	0,0000	0,0000	0,0000	0,0001	0,0004	0,0016	0,0052	0,0139
	13	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001	0,0003	0,0010	0,0032
	14	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001	0,0005
	15	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000

Tabela 5 Probabilidades Binomiais (continuação)

		p								
n	*x*	0,10	0,15	0,20	0,25	0,30	0,35	0,40	0,45	0,50
18	0	0,1501	0,0536	0,0180	0,0056	0,0016	0,0004	0,0001	0,0000	0,0000
	1	0,3002	0,1704	0,0811	0,0338	0,0126	0,0042	0,0012	0,0003	0,0001
	2	0,2835	0,2556	0,1723	0,0958	0,0458	0,0190	0,0069	0,0022	0,0006
	3	0,1680	0,2406	0,2297	0,1704	0,1046	0,0547	0,0246	0,0095	0,0031
	4	0,0700	0,1592	0,2153	0,2130	0,1681	0,1104	0,0614	0,0291	0,0117
	5	0,0218	0,0787	0,1507	0,1988	0,2017	0,1664	0,1146	0,0666	0,0327
	6	0,0052	0,0301	0,0816	0,1436	0,1873	0,1941	0,1655	0,1181	0,0708
	7	0,0010	0,0091	0,0350	0,0820	0,1376	0,1792	0,1892	0,1657	0,1214
	8	0,0002	0,0022	0,0120	0,0376	0,0811	0,1327	0,1734	0,1864	0,1669
	9	0,0000	0,0004	0,0033	0,0139	0,0386	0,0794	0,1284	0,1694	0,1855
	10	0,0000	0,0001	0,0008	0,0042	0,0149	0,0385	0,0771	0,1248	0,1669
	11	0,0000	0,0000	0,0001	0,0010	0,0046	0,0151	0,0374	0,0742	0,1214
	12	0,0000	0,0000	0,0000	0,0002	0,0012	0,0047	0,0145	0,0354	0,0708
	13	0,0000	0,0000	0,0000	0,0000	0,0002	0,0012	0,0045	0,0134	0,0327
	14	0,0000	0,0000	0,0000	0,0000	0,0000	0,0002	0,0011	0,0039	0,0117
	15	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0002	0,0009	0,0031
	16	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001	0,0006
	17	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001
	18	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
20	0	0,1216	0,0388	0,0115	0,0032	0,0008	0,0002	0,0000	0,0000	0,0000
	1	0,2702	0,1368	0,0576	0,0211	0,0068	0,0020	0,0005	0,0001	0,0000
	2	0,2852	0,2293	0,1369	0,0669	0,0278	0,0100	0,0031	0,0008	0,0002
	3	0,1901	0,2428	0,2054	0,1339	0,0716	0,0323	0,0123	0,0040	0,0011
	4	0,0898	0,1821	0,2182	0,1897	0,1304	0,0738	0,0350	0,0139	0,0046
	5	0,0319	0,1028	0,1746	0,2023	0,1789	0,1272	0,0746	0,0365	0,0148
	6	0,0089	0,0454	0,1091	0,1686	0,1916	0,1712	0,1244	0,0746	0,0370
	7	0,0020	0,0160	0,0545	0,1124	0,1643	0,1844	0,1659	0,1221	0,0739
	8	0,0004	0,0046	0,0222	0,0609	0,1144	0,1614	0,1797	0,1623	0,1201
	9	0,0001	0,0011	0,0074	0,0271	0,0654	0,1158	0,1597	0,1771	0,1602
	10	0,0000	0,0002	0,0020	0,0099	0,0308	0,0686	0,1171	0,1593	0,1762
	11	0,0000	0,0000	0,0005	0,0030	0,0120	0,0336	0,0710	0,1185	0,1602
	12	0,0000	0,0000	0,0001	0,0008	0,0039	0,0136	0,0355	0,0727	0,1201
	13	0,0000	0,0000	0,0000	0,0002	0,0010	0,0045	0,0146	0,0366	0,0739
	14	0,0000	0,0000	0,0000	0,0000	0,0002	0,0012	0,0049	0,0150	0,0370
	15	0,0000	0,0000	0,0000	0,0000	0,0000	0,0003	0,0013	0,0049	0,0148
	16	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0003	0,0013	0,0046
	17	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0002	0,0011
	18	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0002
	19	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
	20	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000

Tabela 6 Valores de $e^{-\mu}$

μ	$e^{-\mu}$	μ	$e^{-\mu}$	μ	$e^{-\mu}$
0,00	1,0000	2,00	0,1353	4,00	0,0183
0,05	0,9512	2,05	0,1287	4,05	0,0174
0,10	0,9048	2,10	0,1225	4,10	0,0166
0,15	0,8607	2,15	0,1165	4,15	0,0158
0,20	0,8187	2,20	0,1108	4,20	0,0150
0,25	0,7788	2,25	0,1054	4,25	0,0143
0,30	0,7408	2,30	0,1003	4,30	0,0136
0,35	0,7047	2,35	0,0954	4,35	0,0129
0,40	0,6703	2,40	0,0907	4,40	0,0123
0,45	0,6376	2,45	0,0863	4,45	0,0117
0,50	0,6065	2,50	0,0821	4,50	0,0111
0,55	0,5769	2,55	0,0781	4,55	0,0106
0,60	0,5488	2,60	0,0743	4,60	0,0101
0,65	0,5220	2,65	0,0707	4,65	0,0096
0,70	0,4966	2,70	0,0672	4,70	0,0091
0,75	0,4724	2,75	0,0639	4,75	0,0087
0,80	0,4493	2,80	0,0608	4,80	0,0082
0,85	0,4274	2,85	0,0578	4,85	0,0078
0,90	0,4066	2,90	0,0550	4,90	0,0074
0,95	0,3867	2,95	0,0523	4,95	0,0071
1,00	0,3679	3,00	0,0498	5,00	0,0067
1,05	0,3499	3,05	0,0474	6,00	0,0025
1,10	0,3329	3,10	0,0450	7,00	0,0009
1,15	0,3166	3,15	0,0429	8,00	0,000335
1,20	0,3012	3,20	0,0408	9,00	0,000123
				10,00	0,000045
1,25	0,2865	3,25	0,0388		
1,30	0,2725	3,30	0,0369		
1,35	0,2592	3,35	0,0351		
1,40	0,2466	3,40	0,0334		
1,45	0,2346	3,45	0,0317		
1,50	0,2231	3,50	0,0302		
1,55	0,2122	3,55	0,0287		
1,60	0,2019	3,60	0,0273		
1,65	0,1920	3,65	0,0260		
1,70	0,1827	3,70	0,0247		
1,75	0,1738	3,75	0,0235		
1,80	0,1653	3,80	0,0224		
1,85	0,1572	3,85	0,0213		
1,90	0,1496	3,90	0,0202		
1,95	0,1423	3,95	0,0193		

Tabela 7 Probabilidades de Poisson

Os registros na tabela fornecem a probabilidade de x ocorrências para um processo de Poisson com uma média μ, Por exemplo, quando $\mu = 2,5$, a probabilidade de quatro ocorrências é 0,1336,

	μ									
x	0,1	0,2	0,3	0,4	0,5	0,6	0,7	0,8	0,9	1,0
0	0,9048	0,8187	0,7408	0,6703	0,6065	0,5488	0,4966	0,4493	0,4066	0,3679
1	0,0905	0,1637	0,2222	0,2681	0,3033	0,3293	0,3476	0,3595	0,3659	0,3679
2	0,0045	0,0164	0,0333	0,0536	0,0758	0,0988	0,1217	0,1438	0,1647	0,1839
3	0,0002	0,0011	0,0033	0,0072	0,0126	0,0198	0,0284	0,0383	0,0494	0,0613
4	0,0000	0,0001	0,0002	0,0007	0,0016	0,0030	0,0050	0,0077	0,0111	0,0153
5	0,0000	0,0000	0,0000	0,0001	0,0002	0,0004	0,0007	0,0012	0,0020	0,0031
6	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001	0,0002	0,0003	0,0005
7	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001

	μ									
x	1,1	1,2	1,3	1,4	1,5	1,6	1,7	1,8	1,9	2,0
0	0,3329	0,3012	0,2725	0,2466	0,2231	0,2019	0,1827	0,1653	0,1496	0,1353
1	0,3662	0,3614	0,3543	0,3452	0,3347	0,3230	0,3106	0,2975	0,2842	0,2707
2	0,2014	0,2169	0,2303	0,2417	0,2510	0,2584	0,2640	0,2678	0,2700	0,2707
3	0,0738	0,0867	0,0998	0,1128	0,1255	0,1378	0,1496	0,1607	0,1710	0,1804
4	0,0203	0,0260	0,0324	0,0395	0,0471	0,0551	0,0636	0,0723	0,0812	0,0902
5	0,0045	0,0062	0,0084	0,0111	0,0141	0,0176	0,0216	0,0260	0,0309	0,0361
6	0,0008	0,0012	0,0018	0,0026	0,0035	0,0047	0,0061	0,0078	0,0098	0,0120
7	0,0001	0,0002	0,0003	0,0005	0,0008	0,0011	0,0015	0,0020	0,0027	0,0034
8	0,0000	0,0000	0,0001	0,0001	0,0001	0,0002	0,0003	0,0005	0,0006	0,0009
9	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001	0,0001	0,0001	0,0002

	μ									
x	2,1	2,2	2,3	2,4	2,5	2,6	2,7	2,8	2,9	3,0
0	0,1225	0,1108	0,1003	0,0907	0,0821	0,0743	0,0672	0,0608	0,0550	0,0498
1	0,2572	0,2438	0,2306	0,2177	0,2052	0,1931	0,1815	0,1703	0,1596	0,1494
2	0,2700	0,2681	0,2652	0,2613	0,2565	0,2510	0,2450	0,2384	0,2314	0,2240
3	0,1890	0,1966	0,2033	0,2090	0,2138	0,2176	0,2205	0,2225	0,2237	0,2240
4	0,0992	0,1082	0,1169	0,1254	0,1336	0,1414	0,1488	0,1557	0,1622	0,1680
5	0,0417	0,0476	0,0538	0,0602	0,0668	0,0735	0,0804	0,0872	0,0940	0,1008
6	0,0146	0,0174	0,0206	0,0241	0,0278	0,0319	0,0362	0,0407	0,0455	0,0504
7	0,0044	0,0055	0,0068	0,0083	0,0099	0,0118	0,0139	0,0163	0,0188	0,0216
8	0,0011	0,0015	0,0019	0,0025	0,0031	0,0038	0,0047	0,0057	0,0068	0,0081
9	0,0003	0,0004	0,0005	0,0007	0,0009	0,0011	0,0014	0,0018	0,0022	0,0027
10	0,0001	0,0001	0,0001	0,0002	0,0002	0,0003	0,0004	0,0005	0,0006	0,0008
11	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001	0,0001	0,0001	0,0002	0,0002
12	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001

Tabela 7 Probabilidades de Poisson (continuação)

x	μ 3,1	3,2	3,3	3,4	3,5	3,6	3,7	3,8	3,9	4,0
0	0,0450	0,0408	0,0369	0,0344	0,0302	0,0273	0,0247	0,0224	0,0202	0,0183
1	0,1397	0,1304	0,1217	0,1135	0,1057	0,0984	0,0915	0,0850	0,0789	0,0733
2	0,2165	0,2087	0,2008	0,1929	0,1850	0,1771	0,1692	0,1615	0,1539	0,1465
3	0,2237	0,2226	0,2209	0,2186	0,2158	0,2125	0,2087	0,2046	0,2001	0,1954
4	0,1734	0,1781	0,1823	0,1858	0,1888	0,1912	0,1931	0,1944	0,1951	0,1954
5	0,1075	0,1140	0,1203	0,1264	0,1322	0,1377	0,1429	0,1477	0,1522	0,1563
6	0,0555	0,0608	0,0662	0,0716	0,0771	0,0826	0,0881	0,0936	0,0989	0,1042
7	0,0246	0,0278	0,0312	0,0348	0,0385	0,0425	0,0466	0,0508	0,0551	0,0595
8	0,0095	0,0111	0,0129	0,0148	0,0169	0,0191	0,0215	0,0241	0,0269	0,0298
9	0,0033	0,0040	0,0047	0,0056	0,0066	0,0076	0,0089	0,0102	0,0116	0,0132
10	0,0010	0,0013	0,0016	0,0019	0,0023	0,0028	0,0033	0,0039	0,0045	0,0053
11	0,0003	0,0004	0,0005	0,0006	0,0007	0,0009	0,0011	0,0013	0,0016	0,0019
12	0,0001	0,0001	0,0001	0,0002	0,0002	0,0003	0,0003	0,0004	0,0005	0,0006
13	0,0000	0,0000	0,0000	0,0000	0,0001	0,0001	0,0001	0,0001	0,0002	0,0002
14	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001

x	μ 4,1	4,2	4,3	4,4	4,5	4,6	4,7	4,8	4,9	5,0
0	0,0166	0,0150	0,0136	0,0123	0,0111	0,0101	0,0091	0,0082	0,0074	0,0067
1	0,0679	0,0630	0,0583	0,0540	0,0500	0,0462	0,0427	0,0395	0,0365	0,0337
2	0,1393	0,1323	0,1254	0,1188	0,1125	0,1063	0,1005	0,0948	0,0894	0,0842
3	0,1904	0,1852	0,1798	0,1743	0,1687	0,1631	0,1574	0,1517	0,1460	0,1404
4	0,1951	0,1944	0,1933	0,1917	0,1898	0,1875	0,1849	0,1820	0,1789	0,1755
5	0,1600	0,1633	0,1662	0,1687	0,1708	0,1725	0,1738	0,1747	0,1753	0,1755
6	0,1093	0,1143	0,1191	0,1237	0,1281	0,1323	0,1362	0,1398	0,1432	0,1462
7	0,0640	0,0686	0,0732	0,0778	0,0824	0,0869	0,0914	0,0959	0,1002	0,1044
8	0,0328	0,0360	0,0393	0,0428	0,0463	0,0500	0,0537	0,0575	0,0614	0,0653
9	0,0150	0,0168	0,0188	0,0209	0,0232	0,0255	0,0280	0,0307	0,0334	0,0363
10	0,0061	0,0071	0,0081	0,0092	0,0104	0,0118	0,0132	0,0147	0,0164	0,0181
11	0,0023	0,0027	0,0032	0,0037	0,0043	0,0049	0,0056	0,0064	0,0073	0,0082
12	0,0008	0,0009	0,0011	0,0014	0,0016	0,0019	0,0022	0,0026	0,0030	0,0034
13	0,0002	0,0003	0,0004	0,0005	0,0006	0,0007	0,0008	0,0009	0,0011	0,0013
14	0,0001	0,0001	0,0001	0,0001	0,0002	0,0002	0,0003	0,0003	0,0004	0,0005
15	0,0000	0,0000	0,0000	0,0000	0,0001	0,0001	0,0001	0,0001	0,0001	0,0002

Tabela 7 Probabilidades de Poisson (continuação)

	μ									
x	5,1	5,2	5,3	5,4	5,5	5,6	5,7	5,8	5,9	6,0
0	0,0061	0,0055	0,0050	0,0045	0,0041	0,0037	0,0033	0,0030	0,0027	0,0025
1	0,0311	0,0287	0,0265	0,0244	0,0225	0,0207	0,0191	0,0176	0,0162	0,0149
2	0,0793	0,0746	0,0701	0,0659	0,0618	0,0580	0,0544	0,0509	0,0477	0,0446
3	0,1348	0,1293	0,1239	0,1185	0,1133	0,1082	0,1033	0,0985	0,0938	0,0892
4	0,1719	0,1681	0,1641	0,1600	0,1558	0,1515	0,1472	0,1428	0,1383	0,1339
5	0,1753	0,1748	0,1740	0,1728	0,1714	0,1697	0,1678	0,1656	0,1632	0,1606
6	0,1490	0,1515	0,1537	0,1555	0,1571	0,1587	0,1594	0,1601	0,1605	0,1606
7	0,1086	0,1125	0,1163	0,1200	0,1234	0,1267	0,1298	0,1326	0,1353	0,1377
8	0,0692	0,0731	0,0771	0,0810	0,0849	0,0887	0,0925	0,0962	0,0998	0,1033
9	0,0392	0,0423	0,0454	0,0486	0,0519	0,0552	0,0586	0,0620	0,0654	0,0688
10	0,0200	0,0220	0,0241	0,0262	0,0285	0,0309	0,0334	0,0359	0,0386	0,0413
11	0,0093	0,0104	0,0116	0,0129	0,0143	0,0157	0,0173	0,0190	0,0207	0,0225
12	0,0039	0,0045	0,0051	0,0058	0,0065	0,0073	0,0082	0,0092	0,0102	0,0113
13	0,0015	0,0018	0,0021	0,0024	0,0028	0,0032	0,0036	0,0041	0,0046	0,0052
14	0,0006	0,0007	0,0008	0,0009	0,0011	0,0013	0,0015	0,0017	0,0019	0,0022
15	0,0002	0,0002	0,0003	0,0003	0,0004	0,0005	0,0006	0,0007	0,0008	0,0009
16	0,0001	0,0001	0,0001	0,0001	0,0001	0,0002	0,0002	0,0002	0,0003	0,0003
17	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001	0,0001	0,0001	0,0001	0,0001

	μ									
x	6,1	6,2	6,3	6,4	6,5	6,6	6,7	6,8	6,9	7,0
0	0,0022	0,0020	0,0018	0,0017	0,0015	0,0014	0,0012	0,0011	0,0010	0,0009
1	0,0137	0,0126	0,0116	0,0106	0,0098	0,0090	0,0082	0,0076	0,0070	0,0064
2	0,0417	0,0390	0,0364	0,0340	0,0318	0,0296	0,0276	0,0258	0,0240	0,0223
3	0,0848	0,0806	0,0765	0,0726	0,0688	0,0652	0,0617	0,0584	0,0552	0,0521
4	0,1294	0,1249	0,1205	0,1162	0,1118	0,1076	0,1034	0,0992	0,0952	0,0912
5	0,1579	0,1549	0,1519	0,1487	0,1454	0,1420	0,1385	0,1349	0,1314	0,1277
6	0,1605	0,1601	0,1595	0,1586	0,1575	0,1562	0,1546	0,1529	0,1511	0,1490
7	0,1399	0,1418	0,1435	0,1450	0,1462	0,1472	0,1480	0,1486	0,1489	0,1490
8	0,1066	0,1099	0,1130	0,1160	0,1188	0,1215	0,1240	0,1263	0,1284	0,1304
9	0,0723	0,0757	0,0791	0,0825	0,0858	0,0891	0,0923	0,0954	0,0985	0,1014
10	0,0441	0,0469	0,0498	0,0528	0,0558	0,0588	0,0618	0,0649	0,0679	0,0710
11	0,0245	0,0265	0,0285	0,0307	0,0330	0,0353	0,0377	0,0401	0,0426	0,0452
12	0,0124	0,0137	0,0150	0,0164	0,0179	0,0194	0,0210	0,0227	0,0245	0,0264
13	0,0058	0,0065	0,0073	0,0081	0,0089	0,0098	0,0108	0,0119	0,0130	0,0142
14	0,0025	0,0029	0,0033	0,0037	0,0041	0,0046	0,0052	0,0058	0,0064	0,0071
15	0,0010	0,0012	0,0014	0,0016	0,0018	0,0020	0,0023	0,0026	0,0029	0,0033
16	0,0004	0,0005	0,0005	0,0006	0,0007	0,0008	0,0010	0,0011	0,0013	0,0014
17	0,0001	0,0002	0,0002	0,0002	0,0003	0,0003	0,0004	0,0004	0,0005	0,0006
18	0,0000	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0002	0,0002	0,0002
19	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001	0,0001	0,0001

Tabela 7 Probabilidades de Poisson (continuação)

	μ									
x	7,1	7,2	7,3	7,4	7,5	7,6	7,7	7,8	7,9	8,0
0	0,0008	0,0007	0,0007	0,0006	0,0006	0,0005	0,0005	0,0004	0,0004	0,0003
1	0,0059	0,0054	0,0049	0,0045	0,0041	0,0038	0,0035	0,0032	0,0029	0,0027
2	0,0208	0,0194	0,0180	0,0167	0,0156	0,0145	0,0134	0,0125	0,0116	0,0107
3	0,0492	0,0464	0,0438	0,0413	0,0389	0,0366	0,0345	0,0324	0,0305	0,0286
4	0,0874	0,0836	0,0799	0,0764	0,0729	0,0696	0,0663	0,0632	0,0602	0,0573
5	0,1241	0,1204	0,1167	0,1130	0,1094	0,1057	0,1021	0,0986	0,0951	0,0916
6	0,1468	0,1445	0,1420	0,1394	0,1367	0,1339	0,1311	0,1282	0,1252	0,1221
7	0,1489	0,1486	0,1481	0,1474	0,1465	0,1454	0,1442	0,1428	0,1413	0,1396
8	0,1321	0,1337	0,1351	0,1363	0,1373	0,1382	0,1388	0,1392	0,1395	0,1396
9	0,1042	0,1070	0,1096	0,1121	0,1144	0,1167	0,1187	0,1207	0,1224	0,1241
10	0,0740	0,0770	0,0800	0,0829	0,0858	0,0887	0,0914	0,0941	0,0967	0,0993
11	0,0478	0,0504	0,0531	0,0558	0,0585	0,0613	0,0640	0,0667	0,0695	0,0722
12	0,0283	0,0303	0,0323	0,0344	0,0366	0,0388	0,0411	0,0434	0,0457	0,0481
13	0,0154	0,0168	0,0181	0,0196	0,0211	0,0227	0,0243	0,0260	0,0278	0,0296
14	0,0078	0,0086	0,0095	0,0104	0,0113	0,0123	0,0134	0,0145	0,0157	0,0169
15	0,0037	0,0041	0,0046	0,0051	0,0057	0,0062	0,0069	0,0075	0,0083	0,0090
16	0,0016	0,0019	0,0021	0,0024	0,0026	0,0030	0,0033	0,0037	0,0041	0,0045
17	0,0007	0,0008	0,0009	0,0010	0,0012	0,0013	0,0015	0,0017	0,0019	0,0021
18	0,0003	0,0003	0,0004	0,0004	0,0005	0,0006	0,0006	0,0007	0,0008	0,0009
19	0,0001	0,0001	0,0001	0,0002	0,0002	0,0002	0,0003	0,0003	0,0003	0,0004
20	0,0000	0,0000	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0002
21	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001	0,0001

	μ									
x	8,1	8,2	8,3	8,4	8,5	8,6	8,7	8,8	8,9	9,0
0	0,0003	0,0003	0,0002	0,0002	0,0002	0,0002	0,0002	0,0002	0,0001	0,0001
1	0,0025	0,0023	0,0021	0,0019	0,0017	0,0016	0,0014	0,0013	0,0012	0,0011
2	0,0100	0,0092	0,0086	0,0079	0,0074	0,0068	0,0063	0,0058	0,0054	0,0050
3	0,0269	0,0252	0,0237	0,0222	0,0208	0,0195	0,0183	0,0171	0,0160	0,0150
4	0,0544	0,0517	0,0491	0,0466	0,0443	0,0420	0,0398	0,0377	0,0357	0,0337
5	0,0882	0,0849	0,0816	0,0784	0,0752	0,0722	0,0692	0,0663	0,0635	0,0607
6	0,1191	0,1160	0,1128	0,1097	0,1066	0,1034	0,1003	0,0972	0,0941	0,0911
7	0,1378	0,1358	0,1338	0,1317	0,1294	0,1271	0,1247	0,1222	0,1197	0,1171
8	0,1395	0,1392	0,1388	0,1382	0,1375	0,1366	0,1356	0,1344	0,1332	0,1318
9	0,1256	0,1269	0,1280	0,1290	0,1299	0,1306	0,1311	0,1315	0,1317	0,1318
10	0,1017	0,1040	0,1063	0,1084	0,1104	0,1123	0,1140	0,1157	0,1172	0,1186
11	0,0749	0,0776	0,0802	0,0828	0,0853	0,0878	0,0902	0,0925	0,0948	0,0970
12	0,0505	0,0530	0,0555	0,0579	0,0604	0,0629	0,0654	0,0679	0,0703	0,0728
13	0,0315	0,0334	0,0354	0,0374	0,0395	0,0416	0,0438	0,0459	0,0481	0,0504
14	0,0182	0,0196	0,0210	0,0225	0,0240	0,0256	0,0272	0,0289	0,0306	0,0324
15	0,0098	0,0107	0,0116	0,0126	0,0136	0,0147	0,0158	0,0169	0,0182	0,1094
16	0,0050	0,0055	0,0060	0,0066	0,0072	0,0079	0,0086	0,0093	0,0101	0,0109
17	0,0024	0,0026	0,0029	0,0033	0,0036	0,0040	0,0044	0,0048	0,0053	0,0058
18	0,0011	0,0012	0,0014	0,0015	0,0017	0,0019	0,0021	0,0024	0,0026	0,0029
19	0,0005	0,0005	0,0006	0,0007	0,0008	0,0009	0,0010	0,0011	0,0012	0,0014
20	0,0002	0,0002	0,0002	0,0003	0,0003	0,0004	0,0004	0,0005	0,0005	0,0006
21	0,0001	0,0001	0,0001	0,0001	0,0001	0,0002	0,0002	0,0002	0,0002	0,0003
22	0,0000	0,0000	0,0000	0,0000	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001

Tabela 7 Probabilidades de Poisson (continuação)

x	μ 9,1	9,2	9,3	9,4	9,5	9,6	9,7	9,8	9,9	10
0	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0000
1	0,0010	0,0009	0,0009	0,0008	0,0007	0,0007	0,0006	0,0005	0,0005	0,0005
2	0,0046	0,0043	0,0040	0,0037	0,0034	0,0031	0,0029	0,0027	0,0025	0,0023
3	0,0140	0,0131	0,0123	0,0115	0,0107	0,0100	0,0093	0,0087	0,0081	0,0076
4	0,0319	0,0302	0,0285	0,0269	0,0254	0,0240	0,0226	0,0213	0,0201	0,0189
5	0,0581	0,0555	0,0530	0,0506	0,0483	0,0460	0,0439	0,0418	0,0398	0,0378
6	0,0881	0,0851	0,0822	0,0793	0,0764	0,0736	0,0709	0,0682	0,0656	0,0631
7	0,1145	0,1118	0,1091	0,1064	0,1037	0,1010	0,0982	0,0955	0,0928	0,0901
8	0,1302	0,1286	0,1269	0,1251	0,1232	0,1212	0,1191	0,1170	0,1148	0,1126
9	0,1317	0,1315	0,1311	0,1306	0,1300	0,1293	0,1284	0,1274	0,1263	0,1251
10	0,1198	0,1210	0,1219	0,1228	0,1235	0,1241	0,1245	0,1249	0,1250	0,1251
11	0,0991	0,1012	0,1031	0,1049	0,1067	0,1083	0,1098	0,1112	0,1125	0,1137
12	0,0752	0,0776	0,0799	0,0822	0,0844	0,0866	0,0888	0,0908	0,0928	0,0948
13	0,0526	0,0549	0,0572	0,0594	0,0617	0,0640	0,0662	0,0685	0,0707	0,0729
14	0,0342	0,0361	0,0380	0,0399	0,0419	0,0439	0,0459	0,0479	0,0500	0,0521
15	0,0208	0,0221	0,0235	0,0250	0,0265	0,0281	0,0297	0,0313	0,0330	0,0347
16	0,0118	0,0127	0,0137	0,0147	0,0157	0,0168	0,0180	0,0192	0,0204	0,0217
17	0,0063	0,0069	0,0075	0,0081	0,0088	0,0095	0,0103	0,0111	0,0119	0,0128
18	0,0032	0,0035	0,0039	0,0042	0,0046	0,0051	0,0055	0,0060	0,0065	0,0071
19	0,0015	0,0017	0,0019	0,0021	0,0023	0,0026	0,0028	0,0031	0,0034	0,0037
20	0,0007	0,0008	0,0009	0,0010	0,0011	0,0012	0,0014	0,0015	0,0017	0,0019
21	0,0003	0,0003	0,0004	0,0004	0,0005	0,0006	0,0006	0,0007	0,0008	0,0009
22	0,0001	0,0001	0,0002	0,0002	0,0002	0,0002	0,0003	0,0003	0,0004	0,0004
23	0,0000	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0002	0,0002
24	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001	0,0001	0,0001

x	μ 11	12	13	14	15	16	17	18	19	20
0	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
1	0,0002	0,0001	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
2	0,0010	0,0004	0,0002	0,0001	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
3	0,0037	0,0018	0,0008	0,0004	0,0002	0,0001	0,0000	0,0000	0,0000	0,0000
4	0,0102	0,0053	0,0027	0,0013	0,0006	0,0003	0,0001	0,0001	0,0000	0,0000
5	0,0224	0,0127	0,0070	0,0037	0,0019	0,0010	0,0005	0,0002	0,0001	0,0001
6	0,0411	0,0255	0,0152	0,0087	0,0048	0,0026	0,0014	0,0007	0,0004	0,0002
7	0,0646	0,0437	0,0281	0,0174	0,0104	0,0060	0,0034	0,0018	0,0010	0,0005
8	0,0888	0,0655	0,0457	0,0304	0,0194	0,0120	0,0072	0,0042	0,0024	0,0013
9	0,1085	0,0874	0,0661	0,0473	0,0324	0,0213	0,0135	0,0083	0,0050	0,0029
10	0,1194	0,1048	0,0859	0,0663	0,0486	0,0341	0,0230	0,0150	0,0095	0,0058
11	0,1194	0,1144	0,1015	0,0844	0,0663	0,0496	0,0355	0,0245	0,0164	0,0106
12	0,1094	0,1144	0,1099	0,0984	0,0829	0,0661	0,0504	0,0368	0,0259	0,0176
13	0,0926	0,1056	0,1099	0,1060	0,0956	0,0814	0,0658	0,0509	0,0378	0,0271
14	0,0728	0,0905	0,1021	0,1060	0,1024	0,0930	0,0800	0,0655	0,0514	0,0387

Tabela 7 Probabilidades de Poisson (continuação)

	μ									
x	11	12	13	14	15	16	17	18	19	20
15	0,0534	0,0724	0,0885	0,0989	0,1024	0,0992	0,0906	0,0786	0,0650	0,0516
16	0,0367	0,0543	0,0719	0,0866	0,0960	0,0992	0,0963	0,0884	0,0772	0,0646
17	0,0237	0,0383	0,0550	0,0713	0,0847	0,0934	0,0963	0,0936	0,0863	0,0760
18	0,0145	0,0256	0,0397	0,0554	0,0706	0,0830	0,0909	0,0936	0,0911	0,0844
19	0,0084	0,0161	0,0272	0,0409	0,0557	0,0699	0,0814	0,0887	0,0911	0,0888
20	0,0046	0,0097	0,0177	0,0286	0,0418	0,0559	0,0692	0,0798	0,0866	0,0888
21	0,0024	0,0055	0,0109	0,0191	0,0299	0,0426	0,0560	0,0684	0,0783	0,0846
22	0,0012	0,0030	0,0065	0,0121	0,0204	0,0310	0,0433	0,0560	0,0676	0,0769
23	0,0006	0,0016	0,0037	0,0074	0,0133	0,0216	0,0320	0,0438	0,0559	0,0669
24	0,0003	0,0008	0,0020	0,0043	0,0083	0,0144	0,0226	0,0328	0,0442	0,0557
25	0,0001	0,0004	0,0010	0,0024	0,0050	0,0092	0,0154	0,0237	0,0336	0,0446
26	0,0000	0,0002	0,0005	0,0013	0,0029	0,0057	0,0101	0,0164	0,0246	0,0343
27	0,0000	0,0001	0,0002	0,0007	0,0016	0,0034	0,0063	0,0109	0,0173	0,0254
28	0,0000	0,0000	0,0001	0,0003	0,0009	0,0019	0,0038	0,0070	0,0117	0,0181
29	0,0000	0,0000	0,0001	0,0002	0,0004	0,0011	0,0023	0,0044	0,0077	0,0125
30	0,0000	0,0000	0,0000	0,0001	0,0002	0,0006	0,0013	0,0026	0,0049	0,0083
31	0,0000	0,0000	0,0000	0,0000	0,0001	0,0003	0,0007	0,0015	0,0030	0,0054
32	0,0000	0,0000	0,0000	0,0000	0,0001	0,0001	0,0004	0,0009	0,0018	0,0034
33	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001	0,0002	0,0005	0,0010	0,0020
34	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001	0,0002	0,0006	0,0012
35	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001	0,0003	0,0007
36	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001	0,0002	0,0004
37	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001	0,0002
38	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001
39	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001

Notação de Somatório

Somatório

Definição

$$\sum_{i=1}^{n} x_i = x_1 + x_2 + \cdots + x_n \qquad \text{(C.1)}$$

Exemplo para $x_1 = 5, x_2 = 8, x_3 = 14$:

$$\begin{aligned} \sum_{i=1}^{3} x_i &= x_1 + x_2 + x_3 \\ &= 5 + 8 + 14 \\ &= 27 \end{aligned}$$

Resultado 1

Para uma constante c:

$$\sum_{i=1}^{n} c = (\underbrace{c + c + \cdots + c}_{n \text{ times}}) = nc \qquad \text{(C.2)}$$

Exemplo para $c = 5, n = 10$:

$$\sum_{i=1}^{10} 5 = 10(5) = 50$$

Exemplo para $c = \bar{x}$:

$$\sum_{i=1}^{n} \bar{x} = n\bar{x}$$

Resultado 2

$$\sum_{i=1}^{n} cx_i = cx_1 + cx_2 + \cdots + cx_n$$

$$= c(x_1 + x_2 + \cdots + x_n) = c\sum_{i=1}^{n} x_i \qquad \text{(C.3)}$$

Exemplo para $x_1 = 5, x_2 = 8, x_3 = 14, c = 2$:

$$\sum_{i=1}^{3} 2x_i = 2\sum_{i=1}^{3} x_i = 2(27) = 54$$

Resultado 3

$$\sum_{i=1}^{n} (ax_i + by_i) = a\sum_{i=1}^{n} x_i + b\sum_{i=1}^{n} y_i \qquad \text{(C.4)}$$

Exemplo para $x_1 = 5, x_2 = 8, x_3 = 14, a = 2, y_1 = 7, y_2 = 3, y_3 = 8, b = 4$:

$$\begin{aligned}\sum_{i=1}^{3} (2x_i + 4y_i) &= 2\sum_{i=1}^{3} x_i + 4\sum_{i=1}^{3} y_i \\ &= 2(27) + 4(18) \\ &= 54 + 72 \\ &= 126\end{aligned}$$

Somatórios Duplos

Considere os dados seguintes envolvendo a variável x_{ij}, em que i é o subscrito denotando a posição de linha, e j é o subscrito denotando a posição de coluna:

		Coluna		
		1	2	3
Linha	1	$x_{11} = 10$	$x_{12} = 8$	$x_{13} = 6$
	2	$x_{21} = 7$	$x_{22} = 4$	$x_{23} = 12$

Definição

$$\begin{aligned}\sum_{i=1}^{n}\sum_{j=1}^{m} x_{ij} = &(x_{11} + x_{12} + \cdots + x_{1m}) + (x_{21} + x_{22} + \cdots + x_{2m}) \\ &+ (x_{31} + x_{32} + \cdots + x_{3m}) + \cdots + (x_{n1} + x_{n2} + \cdots + x_{nm})\end{aligned} \qquad \text{(C.5)}$$

Exemplo:

$$\begin{aligned}\sum_{i=1}^{2}\sum_{j=1}^{3} x_{ij} &= x_{11} + x_{12} + x_{13} + x_{21} + x_{22} + x_{23} \\ &= 10 + 8 + 6 + 7 + 4 + 12 \\ &= 47\end{aligned}$$

Definição

$$\sum_{i=1}^{n} x_{ij} = x_{1j} + x_{2j} + \cdots + x_{nj} \qquad \text{(C.6)}$$

Exemplo:

$$\begin{aligned}\sum_{i=1}^{2} x_{i2} &= x_{12} + x_{22} \\ &= 8 + 4 \\ &= 12\end{aligned}$$

Notação Simplificada

Algumas vezes, quando um somatório se refere a todos os valores do subscrito, usamos as seguintes notações simplificadas:

$$\sum_{i=1}^{n} x_i = \sum x_i \tag{C.7}$$

$$\sum_{i=1}^{n}\sum_{j=1}^{m} x_{ij} = \sum\sum x_{ij} \tag{C.8}$$

$$\sum_{i=1}^{n} x_{ij} = \sum_{i} x_{ij} \tag{C.9}$$

Soluções dos Autotestes e Respostas dos Exercícios Pares

Capítulo 1

2. **a.** 9
b. 4
c. Qualitativa: país e preço do quarto
Quantitativa: número de quartos e pontuação global
d. O país é nominal; o preço dos quartos é ordinal; o número dos quartos é uma razão; a pontuação global é um intervalo

3. **a.** Número médio dos quartos = 808/9 = 89,78, ou aproximadamente 90 quartos
b. Pontuação global = 732,1/9 = 81,3
c. Dois dos nove estão localizados na Inglaterra; aproximadamente 22%
d. Quatro dos nove têm preços de quartos iguais a US$; aproximadamente 44%

4. **a.** 10
b. Todas as marcas de *minisystems* manufaturados
c. US$ 314,00
d. US$ 314,00

6. As perguntas a, c e d fornecem dados quantitativos
As perguntas b e e fornecem dados qualitativos

8. **a.** 1.005
b. Qualitativos
c. Porcentagens
d. Aproximadamente 291

10. **a.** Quantitativos; razão
b. Qualitativos; nominal
c. Qualitativos; ordinal
d. Quantitativos; razão
e. Qualitativos; nominal

12. **a.** Todos os que visitam o Havaí
b. Sim
c. A primeira e a quarta perguntas fornecem dados quantitativos
A segunda e a terceira perguntas fornecem dados qualitativos

13. **a.** Quantitativos
b. Série histórica com seis observações
c. Os ganhos da Volkswagen
d. Seria esperado um aumento em 2003, mas parece que a taxa de crescimento está se desacelerando

14. **a.** Qualitativa

16. **a.** Testes de sabor do produto e teste de marketing
b. Com estudos estatísticos especialmente projetados

18. **a.** 36%
b. 189
c. Qualitativos

20. **a.** 43% dos gerentes eram especuladores otimistas (*bullish*) e 21% dos gerentes esperavam que setor da saúde ocupasse a posição de liderança na indústria ao longo dos 12 meses.
b. A estimativa do rendimento médio em 12 meses é 11,2% para a população de gerentes de investimentos.
c. A média amostral de 2,5 anos é uma estimativa de quanto tempo a população de gerentes de investimento acha que será necessário para retomar o crescimento sustentável.

22. **a.** Todos os eleitores registrados na Califórnia
b. Os eleitores registrados contatados pelo Policy Institute
c. Porque consome muito tempo e é muito custoso envolver a população inteira

24. **a.** Correta
b. Incorreta
c. Correta
d. Incorreta
e. Incorreta

Capítulo 2

2. **a** 0,20

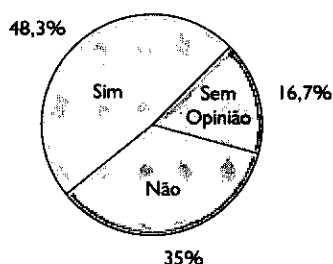
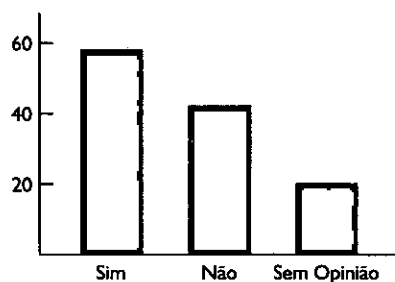
b. 40

c/d.

Classe	Freqüência	Freqüência Percentual
A	44	22
B	36	18
C	80	40
D	40	20
Total	200	100

3. **a.** 360º × 58/120 = 174º

b. 360º × 42/120 = 126º

c.

d.

4. **a.** Qualitativos

b.

Programa de TV	Freqüência	Freqüência Percentual
CSI	18	36
ER	11	22
Friends	15	30
Raymond	6	12
Total	50	100

d. CSI teve a maior audiência; Friends ficou em segundo lugar.

6. **a.**

Livro	Freqüência	Freqüência Percentual
7 Habits	10	16,66
Millionaire	16	26,67
Motley	9	15,00
Dad	13	21,67
WSJ Guide	6	10,00
Outros	6	10,00
Total	60	100,00

b. Os cinco primeiros colocados: *Millionaire, Dad, 7 Habits, Motley, WSJ Guide*

c. 48,33%

7.

Avaliação	Freqüência	Freqüência Percentual
Excelente	19	0,38
Ótimo	13	0,26
Bom	10	0,20
Médio	6	0,12
Fraco	2	0,04

A administração deve estar satisfeita com estes resultados: 64% das avaliações variaram de ótimo a excelente, e 84% das avaliações são boas ou melhores; comparar essas avaliações com os resultados anteriores mostrará se as avaliações dos clientes estão melhorando a qualidade das refeições.

8. **a.**

Posição	Freqüência	Freqüência Percentual
Arremessador (*Pitcher*) – A	17	0,309
Receptor (*catcher*) – R	4	0,73
Primeira base (1)	5	0,091
Segunda base (2)	4	0,073
Terceira base (3)	2	0,036
Interbase (*shortstop*)	5	0,091
Jardineiro esquerdo (E)	6	0,109
Jardineiro Central (C)	5	0,091
Jardineiro Direito (D)	7	0,127
Totais	55	1,000

b. Arremessador (*pitcher*)

c. Terceira base (3)

d. Jardineiro direito

e. 16 *infielders* (1, 2, 3 e I) para 18 *outfielders* (E, C e D)

10. **a.** Os dados são ordinais; eles simplesmente fornecem classificações de acordo com a qualidade.

b.

Resposta	Freqüência	Freqüência Percentual
3	2	0,03
4	4	0,07
5	12	0,20
6	24	0,40
7	18	0,30
Totais	18	30

12.

Classe	Freqüência Cumulativa	Freqüência Relativa Cumulativa
≤19	10	0,20
≤29	24	0,48
≤39	41	0,82
≤49	48	0,96
≤59	50	1,00

14. b/c.

Classe	Freqüência	Freqüência Percentual
6,0 a 7,9	4	20
8,0 a 9,9	2	10
10,0 a 11,9	8	40
12,0 a 13,9	3	15
14,0 a 15,9	3	15
Totais	20	100

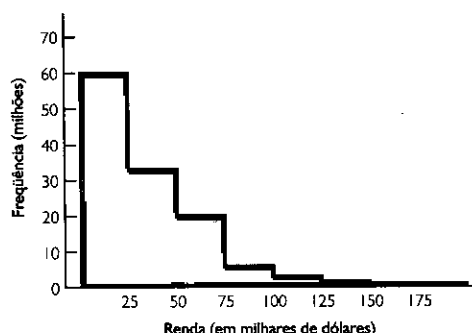
15. a/b.

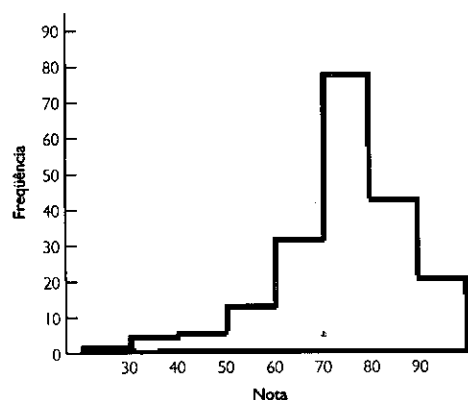
Tempo de Espera	Freqüência	Freqüência Relativa
0 a 4	4	0,20
5 a 9	8	0,40
10 a 14	5	0,25
15 a 19	2	0,10
20 a 24	1	0,05
Totais	20	1,00

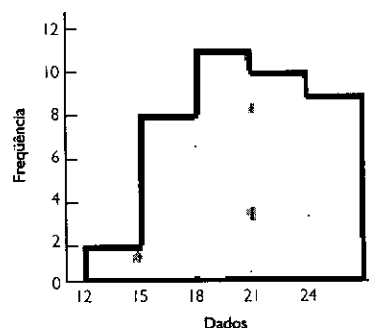
c/d.

Tempo de Espera	Freqüência Cumulativa	Freqüência Relativa Cumulativa
≤4	4	0,20
≤9	12	0,60
≤14	17	0,85
≤19	19	0,95
≤24	20	1,00

e. 12/20 = 0,60

16. a. Renda Bruta Ajustada

O histograma apresenta uma assimetria à direita.

b. Notas de Exame

O histograma apresenta uma assimetria à esquerda.

c.

O histograma apresenta uma leve assimetria à esquerda, mas é aproximadamente simétrico.

18. a. O menor salário: US$ 93 mil
O maior salário: US$ 178 mil

b.

Salário (US$ 1.000)	Freqüência	Freqüência Relativa	Freqüência Percentual
91 a 105	4	0,08	8
106 a 120	5	0,10	10
121 a 135	11	0,22	22
136 a 150	18	0,36	36
151 a 165	9	0,18	18
166 a 180	3	0,06	6
Total	50	1,00	100

c. 20/50
d. 24%

20. a.

Preço	Freqüência	Freqüência Percentual
30 a 39,99	7	35
40 a 49,99	5	25
50 a 59,99	2	10
60 a 69,99	3	15
70 a 79,99	3	15
Total	20	100

c. Fletwood Mac, Harper/Johnson

22.

5 | 7 8
6 | 4 5 8
7 | 0 2 2 5 5 6 8
8 | 0 2 3 5

23. Unidade de folha = 0,1

6 | 3
7 | 5 5 7
8 | 1 3 4 8
9 | 3 6
10 | 0 4 5
11 | 3

24. Unidade de folha = 10

11 | 6
12 | 0 2
13 | 0 6 7
14 | 2 2 7
15 | 5
16 | 0 2 8
17 | 0 2 3

25.

9 8 | 9
10 2 | 4 6 6
11 4 | 5 7 8 8 9
12 2 | 4 5 7
13 1 | 2
14 4 |
15 1 |

26. a.

1 | 0 3 7 7
2 | 4 5 5
3 | 0 0 5 5 9
4 | 0 0 0 5 5 8
5 | 0 0 0 4 5 5

b.

0 | 5 7
1 | 1 1 1 3 4
1 | 5 5 5 8
2 | 0 0 0 0 0 0
2 | 5 5
3 | 0 0 0
3 | 6
4 |
4 |
5 |
5 |
6 | 3

28. a.

2 | 14
2 | 67
3 | 011123
3 | 5677
4 | 003333344
4 | 6679
5 | 00022
5 | 5679
6 | 14
6 | 6
7 | 2

b. 40 a 44 anos, com nove corredores
c. 43 anos, com cinco pessoas
d. 10%; participação relativamente pequena na corrida

29. a.

		y		
		1	2	Total
x	A	5	0	5
	B	11	2	13
	C	2	10	12
	Total	18	12	30

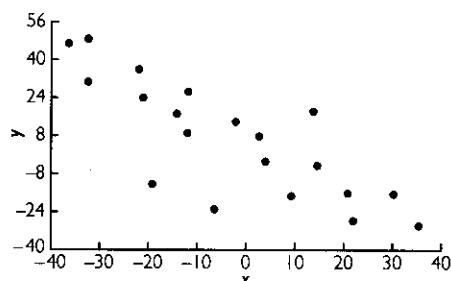
b.

		y		
		1	2	Total
x	A	100,0	0,0	100,0
	B	84,6	15,4	100,0
	C	16,7	83,3	100,0

c.

		y	
		1	2
x	A	27,8	0,0
	B	61,1	16,7
	C	11,1	83,3
	Total	100,0	100,0

d. Os valores de A estão sempre em $y = 1$
Os valores de B estão com maior freqüência em $y = 1$
Os valores de C estão com maior freqüência em $y = 2$

30. a.

b. Há uma relação negativa entre x e y; y decresce à medida que x aumenta.

32. a.

	Renda Familiar (US$1.000)					
Nível Educacional	Menos 25	25.0–49.9	50.0–74.9	75.0–99.9	100 ou mais	Total
Sem Diploma de Ensino Médio	32,70	14,82	8,27	5,02	2,53	15,86
Com Diploma de Ensino Médio	35,74	35,56	31,48	25,39	14,47	30,78
Universitário Incompleto	21,17	29,77	30,25	29,82	22,26	26,37
Grau de Bacharel	7,53	14,43	20,56	25,03	33,88	17,52
Grau Superior a Bacharel	2,86	5,42	9,44	14,74	26,86	9,48
Total	100,00	100,00	100,00	100,00	100,00	100,00

15,86% dos chefes de família não têm diploma de ensino médio.

b. 26,86%, 39,72%

34. a.

Vendas/ Margens/ de Lucro RPL	Avaliação do Lucro por Ação 0–19	20–39	40–59	60–79	80–100	Total
A				1	8	9
B		1	4	5	2	12
C	1		1	2	3	7
D	3	1		1		5
E		2	1			3
Total	4	4	6	9	13	36

b.

Vendas/ Margens/ de Lucro RPL	Avaliação do Lucro por Ação 0–19	20–39	40–59	60–79	80–100	Total
A				11,11	88,89	100
B		8,33	33,33	41,67	16,67	100
C	14,29		14,29	28,57	42,86	100
D	60,00	20,00		20,00		100
E		66,67	33,33			100

Avaliações mais altas do LPA parecem estar associadas a avaliações mais altas das Vendas/Margens de Lucro/RPL.

36. b. Não há relação aparente

38. a.

Veículo	Freqüência	Freqüência Percentual
Accord	6	12
Camry	7	14
F-Series	14	28
Ram	10	20
Silverado	13	26

b. A caminhonete Ford F-Series e o carro de passageiros Toyota Camry

40. a.

Resposta	Freqüência	Freqüência Percentual
Precisão	16	16
Tacadas de aproximação (*approach*)	3	3
Abordagem mental	17	17
Força	8	8
Prática	15	15
Putting (tacada de curto alcance)	10	10
Jogada curta	24	24
Decisões estratégicas	7	7
Total	100	100

b. Má jogada curta, abordagem mental ruim, falta de precisão e prática limitada

42. a/b.

Preço de Fechamento	Freqüência	Freqüência Relativa	Cum. Freq.	Freqüência Relativa Cumulativa
0–9,99	9	0,225	9	0,225
10–19,99	10	0,250	19	0,475
20–29,99	5	0,125	24	0,600
30–39,99	11	0,275	35	0,875
40–49,99	2	0,050	37	0,925
50–59,99	2	0,050	39	0,975
60–69,99	0	0,000	39	0,975
70–79,99	1	0,025	40	1,000
Total	40	1,000		

44.

Renda ($)	Freqüência	Freqüência Relativa
18.000–21.999	13	0,255
22.000–25.999	20	0,392
26.000–29.999	12	0,235
30.000–33.999	4	0,078
34.000–37.999	2	0,039
Total	51	1,000

46. a. Temperatura Máxima

```
3 |
4 |
5 | 7
6 | 1  4  4  4  4  6  8
7 | 3  5  7  9
8 | 0  1  1  4  6
9 | 0  2  3
```

b. Temperatura Mínima

```
3 | 9
4 | 3  6  8
5 | 0  0  0  2  4  4  5  5  7  9
6 | 1  8
7 | 2  4  5  5
8 |
9 |
```

c. A faixa de temperaturas mínimas está abaixo da faixa de temperaturas máximas

d. Oito cidades

e.

Temperatura (°C)	Freqüência Temperatura Máxima	Freqüência Temperatura Mínima
−1,1 a 3,8	0	1
4,4 a 9,4	0	3
10 a 15	1	10
15,5 a 20,5	7	2
21,1 a 26,1	4	4
26,6 a 31,6	5	0
32,2 a 37,2	3	0
Total	20	20

48. a.

Ocupação	Nível de Satisfação 30–39	40–49	50–59	60–69	70–79	80–89	Total
Marceneiro			2	4	3	1	10
Advogado	1	5	2	1	1		10
Fisioterapeuta			5	2	1	2	10
Analista de Sistemas		2	1	4	3		10
Total	1	7	10	11	8	3	40

b.

Ocupação	Nível de Satisfação 30–39	40–49	50–59	60–69	70–79	80–89	Total
Marceneiro			20	40	30	10	100
Advogado	10	50	20	10	10		100
Fisioterapeuta			50	20	10	20	100
Analista de Sistemas		20	10	40	30		100

c. Os marceneiros parecem ter os mais altos níveis de satisfação no trabalho; os advogados, os mais baixos.

50. a. Totais de linha: 247, 54, 82, 121
Totais de coluna: 149, 317, 17, 7, 14

b.

Ano	Freqüência	Combustível	Freqüência
1973 ou antes	247	Eletricidade	149
1974 a 79	54	Gás Natural	317
1980 a 86	82	Petróleo	17
1987 a 91	121	Gás Propano	7
Total	504	Outros	14
		Total	504

c. Tabulação cruzada das porcentagens de coluna

Ano de Construção	Tipo de Combustível				
	Eletricidade	Gás Natural	Petróleo	Gás Propano	Outros
1973 ou antes	26,9	57,7	70,5	71,4	50,0
1974–1979	16,1	8,2	11,8	28,6	0,0
1980–1986	24,8	12,0	5,9	0,0	42,9
1987–1991	32,2	22,1	11,8	0,0	7,1
Total	100,0	100,0	100,0	100,0	100,0

d. Tabulação cruzada das porcentagens de linha

Ano de Construção	Tipo de Combustível					
	Eletricidade	Gás Natural	Petróleo	Gás Propano	Outros	Total
1973 ou antes	16,2	74,1	4,9	2,0	2,8	100,0
1974–1979	44,5	48,1	3,7	3,7	0,0	100,0
1980–1986	45,1	46,4	1,2	0,0	7,3	100,0
1987–1991	39,7	57,8	1,7	0,0	0,8	100,0

52. a. Tabulação cruzada do valor de mercado e lucro

Valor de Mercado (milhares de dólares)	Lucro (milhares de dólares)				
	0–300	300–600	600–900	900–1.200	Total
0–8.000	23	4			27
8.000–16.000	4	4	2	2	12
16.000–24.000		2	1	1	4
24.000–32.000		1	2	1	4
32.000–40.000		2	1		3
Total	27	13	6	4	50

b. Tabulação cruzada das porcentagens de linha

Valor de Mercado (milhares de dólares)	Lucro (milhares de dólares)				
	0–300	300–600	600–900	900–1.200	Total
0–8.000	85,19	14,81	0,00	0,00	100
8.000–16.000	33,33	33,33	16,67	16,67	100
16.000–24.000	0,00	50,00	25,00	25,00	100
24.000–32.000	0,00	25,00	50,00	25,00	100
32.000–40.000	0,00	66,67	33,33	0,00	100

c. Uma relação positiva é indicada entre o lucro e o valor de mercado; à medida que o lucro cresce, o valor de mercado também aumenta

54. b. Uma relação positiva é demonstrada entre o valor de mercado e patrimônio dos acionistas

Capítulo 3

2. 16, 16,5

3. Organize os dados na seguinte ordem: 15, 20, 25, 25, 27, 28, 30, 34

$i = \frac{20}{100}(8) = 1{,}6$; arredonde para cima, para a posição 2

20º percentil = 20

$i = \frac{25}{100}(8) = 2$; use as posições 2 e 3

25º percentil $= \frac{20 + 25}{2} = 22{,}5$

$i = \frac{65}{100}(8) = 5{,}2$; arredonde para cima, para a posição 6

65º percentil = 28

$i = \frac{75}{100}(8) = 6$; use as posições 6 e 7

75º percentil $= \frac{28 + 30}{2} = 29$

4. 59,727, 57, 53

6. a. 422
b. 380
c. 690
d. Não utilizarem a capacidade

8. a. $\bar{x} = \frac{\Sigma x_i}{n} = \frac{695}{20} = 34{,}75$

Moda = 25 (aparece três vezes)

b. Organize os dados na seguinte ordem: 18, 20, 25, 25, 25, 26, 27, 27, 28, 33, 36, 37, 40, 40, 42, 45, 46, 48, 53, 54

Mediana (10ª e 11ª posições)

$\frac{33 + 36}{2} = 34{,}5$

Quem trabalha em casa é ligeiramente mais jovem

c. $i = \frac{25}{100}(20) = 5$; use as posições 5 e 6

$Q_1 = \frac{25 + 26}{2} = 25{,}5$

$i = \frac{75}{100}(20) = 15$; use as posições 15 e 16

$Q_3 = \frac{42 + 45}{2} = 43{,}5$

d. $i = \frac{32}{100}(20) = 6{,}4$; arredonde para a posição 7

32º percentil = 27

No mínimo, 32% das pessoas têm 27 anos ou menos

10. a. 76, 76
b. 39, 37,5
c. Sim; a espera de vagas das salas de emergência é muito longa

12. a. US$ 639,00
b. 98,8 fotografias
c. 110,2 minutos

14. 16,4

15. Intervalo = 34 – 15 = 19

Organize os dados na seguinte ordem: 15, 20, 25, 25, 27, 28, 30, 34

$i = \frac{25}{100}(8) = 2;\ Q_1 = \frac{20 + 25}{2} = 22{,}5$

$i = \frac{75}{100}(8) = 6;\ Q_3 = \frac{28 + 30}{2} = 29$

$\text{IQR} = Q_3 - Q_1 = 29 - 22{,}5 = 6{,}5$

$\bar{x} = \frac{\Sigma x_i}{n} = \frac{204}{8} = 25{,}5$

x_i	$(x_i - x)$	$(x_i - x)^2$
27	1,5	2,25
25	−0,5	0,25
20	−5,5	30,25
15	−10,5	110,25
30	4,5	20,25
34	8,5	72,25
28	2,5	6,25
25	−0,5	0,25

$$s^2 = \frac{\Sigma(x_i - \bar{x})^2}{n - 1} = \frac{242}{8 - 1} = 34,57$$

$$s = \sqrt{34,57} = 5,88$$

16. **a.** Amplitude = 190 − 168 = 22

b. $\bar{x} = \frac{\Sigma x_i}{n} = \frac{1.068}{6} = 178$

$$s^2 = \frac{\Sigma(x_i - \bar{x})^2}{n - 1}$$

$$= \frac{4^2 + (-10)^2 + 6^2 + 12^2 + (-8)^2 + (-4)^2}{6 - 1}$$

$$= \frac{376}{5} = 75,2$$

c. $s = \sqrt{75,2} = 8,67$

d. $\frac{s}{\bar{x}}(100) = \frac{8,67}{178}(100\%) = 4,87\%$

18. **a.** 38; 97; 9,85

b. A região leste apresenta mais variação

20. *Dawson*: amplitude = 2, s = 0,67

Clark: amplitude = 8, s = 2,58

22. **a.** 45,05; 23,98; 57,50; 11,475

b. 190,67; 13,81; 140,63; 11,86

c. 38,02%; 57,97%

d. Maior para os que fazem transações auxiliadas por corretores

24. *Corredores de 400 metros: s* = 0,0564, coeficiente de variação = 5,8%

Meio-fundistas de 1 milha: s = 0,1295, coeficiente de variação = 2,9%

26. 0,20; 1,50; 0; −0,50; −2,20

27. Teorema de Chebyshev: *no mínimo* $(1 - 1/z^2)$

a. $z = \frac{40 - 30}{5} = 2;\ 1 - \frac{1}{(2)^2} = 0,75$

b. $z = \frac{45 - 30}{5} = 3;\ 1 - \frac{1}{(3)^2} = 0,89$

c. $z = \frac{38 - 30}{5} = 1,6;\ 1 - \frac{1}{(1.6)^2} = 0,61$

d. $z = \frac{42 - 30}{5} = 2,4;\ 1 - \frac{1}{(2.4)^2} = 0,83$

e. $z = \frac{48 - 30}{5} = 3,6;\ 1 - \frac{1}{(3.6)^2} = 0,92$

28. **a.** 95%

b. Quase todos

c. 68%

29. **a.** z = 2 desvios padrão

$1 - \frac{1}{z^2} = 1 - \frac{1}{2^2} = \frac{3}{4}$; no mínimo, 75%

b. z = 2,5 desvios padrão

$1 - \frac{1}{z^2} = 1 - \frac{1}{2.5^2} = 0,84$; no mínimo, 84%

c. z = 2 desvios padrão

Regra empírica: 95%

30. **a.** 68%

b. 81,5%

c. 2,5%

32. **a.** −0,67

b. 1,50

c. Nenhuma delas é um ponto fora da curva

d. Sim; z = 8,25

34. **a.** 76,5; 7

b. 16%, 2,5%

c. 12,2; 7,89; não

36. 15; 22,5; 26; 29; 34

38. Organize os dados nesta ordem: 5, 6, 8, 10, 10, 12, 15, 16, 18

$i = \frac{25}{100}(9) = 2.25$; arredonde para a posição 3

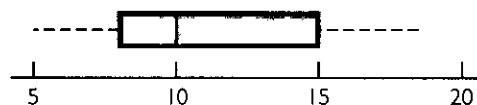
$Q_1 = 8$

Mediana (5ª posição) = 10

$i = \frac{75}{100}(9) = 6,75$; arredonde para a posição 7

$Q_3 = 15$

Regra dos cinco itens: 5, 8, 10, 15, 18

40. **a.** 619, 725, 1.016, 1.699, 4.450

b. Limites: 0, 3.160

c. Sim

d. Não

41. **a.** Organize os dados e ordem crescente

$i = \frac{25}{100}(21) = 5,25$; arredonde para a 6ª posição

$Q_1 = 1.872$

Mediana (11ª posição) = 4.019

$i = \frac{75}{100}(21) = 15,75$; arredonde para a 16ª posição

$Q_3 = 8.305$

Regra dos cinco itens: 608. 1.872. 4.019, 8.305, 14.138

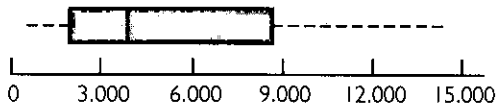
b. AIQ = $Q_3 - Q_1$ = 8.305 − 1.872 = 6.433

Limite inferior: 1.872 − 1,5(6.433) = −7.777

Limite superior: 8.305 + 1,5(6.433) = 17.955

c. Não; os dados estão dentro dos limites

d. 41.138 > 27.604; 41.138 seria um ponto fora da curva; o valor de dados seria revisado e corrigido

e.

42. a. 61
b. 34, 45, 61, 90, 126
c. Não; limite superior – 157,5

44. a. 18,2; 15,35
b. 11,7; 23,5
c. 3,4; 11,7; 15,35; 23,5; 41,3
d. Sim; Alger Small Cap 41,3

45. b. Parece haver uma relação negativa entre x e y

c.

x_i	y_i	$x_i - x$	$y_i - y$	$(x_i - x)(y_i - y)$
4	50	–4	4	–16
6	50	–2	4	–8
11	40	3	–6	–18
3	60	–5	14	–70
16	30	8	–16	–128
40	230	0	0	–240

$$\bar{x} = 8; \bar{y} = 46$$

$$s_{xy} = \frac{\Sigma(x_i - \bar{x})(y_i - \bar{y})}{n-1} = \frac{-240}{4} = -60$$

A covariância da amostra indica uma associação linear negativa entre x e y.

d. $r_{xy} = \dfrac{s_{xy}}{s_x s_y} = \dfrac{-60}{(5{,}43)(11{,}40)} = -0{,}969$

O coeficiente de correlação amostral igual a –0,969 é um indicativo de forte relação linear negativa.

46. b. Parece haver uma relação linear positiva entre x e y
c. $s_{xy} = 26{,}5$
d. $r_{xy} = 0{,}693$

48. –0,91; relação negativa

50. a. 0,92
b. Forte relação linear positiva

52. a. 3,69
b. 3,175

53. a.

f_i	M_i	f_iM_i
4	5	20
7	10	70
9	15	135
5	20	100
25		325

$$\bar{x} = \frac{\Sigma f_i M_i}{n} = \frac{325}{25} = 13$$

b.

f_i	M_i	$(M_i - x)$	$(M_i - x)^2$	$f_i(M_i - x)^2$
4	5	–8	64	256
7	10	–3	9	63
9	15	2	4	36
5	29	7	49	245
25				600

$$s^2 = \frac{\Sigma f_i(M_i - \bar{x})^2}{n-1} = \frac{600}{25-1} = 25$$

$$s = \sqrt{25} = 5$$

54. a.

GPA x_i	Média Ponderada w_i
4 (A)	9
3 (B)	15
2 (C)	33
1 (D)	3
0 (F)	0
	60 horas-crédito

$$\bar{x} = \frac{\Sigma w_i x_i}{\Sigma w_i} = \frac{9(4) + 15(3) + 33(2) + 3(1)}{9 + 15 + 33 + 3} = \frac{150}{60} = 2{,}5$$

b. Sim

56. 10,74; 25,63; 5,06; Estimativa = 1.288,8

58. a. 1.800; 1.351
b. 387; 1.710
c. 7.280; 1.323
d. 3.675.303; 1.917
e. 9.271,01; 96,29; altamente positiva
f. Usando-se um desenho esquemático (*box plot*): 4.135 e 7.450

60. a. 2,3; 1,85
b. 1,90; 1,38
c. Altria Group 5%
d. –51, abaixo da média
e. 1,02, acima da média
f. Não

62. a. $\bar{x} = 83{,}135$; $s = 16{,}173$
b. US$ 50.789 a US$ 115.481
c. A mesma amplitude do item (b); a probabilidade é mais elevada
d. Danbury, CT, é um ponto fora da curva

64. a. 502,67; relação linear positiva
b. 0,933

66. b. 0,9856, forte relação positiva

68. a. 817
b. 833

70. a. 60,68
b. $s^2 = 31{,}23$; $s = 5{,}59$

Capítulo 4

2. $\binom{6}{3} = \frac{6!}{3!3!} = \frac{6\cdot5\cdot4\cdot3\cdot2\cdot1}{(3\cdot2\cdot1)(3\cdot2\cdot1)} = 20$

ABC	ACE	BCD	BEF
ABD	ACF	BCE	CDE
ABE	ADE	BCF	CDF
ABF	ADF	BDE	CEF
ACD	AEF	BDF	DEF

4. **b.** (Cara, Cara, Cara), (Cara, Cara, Coroa), (Cara, Coroa, Cara), (Cara, Coroa, Coroa)
   (Coroa, Cara, Cara), (Coroa, Cara, Coroa), (Coroa, Coroa, Cara), (Coroa, Coroa, Coroa)
   c. $1/8$

6. $P(E_1) = 0,40$, $P(E_2) = 0,26$, $P(E_3) = 0,34$
   Foi utilizado o método de freqüência relativa.

8. **a.** 4: A Comissão faz uma recomendação positiva – A Câmara aprova
   A Comissão faz uma recomendação positiva – A Câmara desaprova
   A Comissão faz uma recomendação negativa – A Câmara aprova
   A Comissão faz uma recomendação negativa – A Câmara desaprova

9. $\binom{50}{4} = \frac{50!}{4!46!} = \frac{50\cdot49\cdot48\cdot47}{4\cdot3\cdot2\cdot1} = 230,300$

10. **a.** Use o critério de freqüência relativa
    P(Califórnia) $= 1.434/2.374 = 0,60$
    b. O número das empresas que não são dos quatro estados:
    $= 2.374 - 1,434 - 390 - 317 - 112$
    $= 221$
    P(Nenhum dos quatro estados) $= 221/2.374 = 0,09$
    c. P(Não estar nas primeiras etapas) $= 1 - 0,22 = 0,78$
    d. Estimativa das empresas de Massachussetts que estão na primeira etapa de desenvolvimento $= (0,22)390 \approx 86$
    e. Se admitirmos que a quantia total dos fundos investidos não difere por estado, podemos multiplicar a probabilidade de um valor destinado ao Colorado pelo total de fundos de investimentos desembolsados para obtermos uma estimativa.
    Estimativa de fundos destinados ao Colorado
    = (112/2.374)(US$ 32,4)
    = US$ 1,53 bilhão
    Nota do autor: A verba real destinada ao Colorado foi de US$ 1,74 bilhão.

12. **a.** 2.869.685
    b. 1/2.869.685
    c. 1/120.526.770

14. **a.** $1/4$
    b. $1/2$
    c. $3/4$

15. **a.** S = (ás de paus, ás de ouro, ás de copas, ás de espadas)
    b. S = (2 de paus, 3 de paus, . . . 10 de paus, J de paus, Q de paus, K de paus, A de paus)
    c. Há 12; valete, rainha ou rei em cada um dos quatro naipes.
    d. Para (a): $4/52 = 1/13 = 0,08$
    Para (b): $13/52 = _ = 0,25$
    Para (c): $12/52 = 0,23$

16. **a.** 36
    c. $1/6$
    d. $5/18$
    e. Não; P(ímpar) $= P$(par) $= \frac{1}{2}$
    f. Clássico

17. **a.** (4, 6), (4, 7), (4, 8)
    b. $0,05 + 0,10 + 0,15 = 0,30$
    c. (2,8), (3,8), (4,8)
    d. $0,05 + 0,05 + 0,15 = 0,25$
    e. **0,15**

18. **a.** $P(0) = 0,05$
    b. P (4 ou 5) $= 0,20$
    c. P(0, 1 ou 20 $= 0,55$

20. **a.** 0,112
    b. 0,086
    c. 0,49

22. **a.** 0,40; 0,40; 0,60
    b. 0,80, sim
    c. $A^C = (E_3, E_4, E_5); C^C = (E_1, E_4)$;
    $P(A^C) = 0,60; P(C^C) = 0,40$
    d. (E_1, E_2, E_5); 0,60
    e. 0,80

23. **a.** $P(A) = P(E_1) + P(E_4) + P(E_6)$
    $= 0,05 + 0,25 + 0,10 = 0,40$
    $P(B) = P(E_2) + P(E_4) + P(E_7)$
    $= 0,20 + 0,25 + 0,05 = 0,50$
    $P(C) = P(E_2) + P(E_3) + P(E_5) + P(E_7)$
    $= 0,20 + 0,20 + 0,15 + 0,05 = 0,60$
    b. $A \cup B = \{E_1, E_2, E_4, E_6, E_7\}$;
    $P(A \cup B) = P(E_1) + P(E_2) + P(E_4) + P(E_6) + P(E_7)$
    $= 0,05 + 0,20 + 0,25 + 0,10 + 0,05$
    $= 0,65$
    c. $A \cap B = \{E_4\}$; $P(A \cap B) = P(E_4) = 0,25$
    d. Sim, eles são mutuamente exclusivos
    e. $B^c = \{E_1, E_3, E_5, E_6\}$;
    $P(B^c) = P(E_1) + P(E_3) + P(E_5) + P(E_6)$
    $= 0,05 + 0,20 + 0,15 + 0,10$
    $= 0,50$

24. **a.** 0,05
    b. 0,70

26. **a.** 0,30; 0,23
    b. 0,17
    c. 0,64

28. Admitamos que B = um assinante alugou um carro por razões comerciais

P = um assinante alugou um carro por razões pessoais

a. $P(B \cup P) = P(B) + P(P) - P(B \ll P)$
$= 0{,}540 + 0{,}458 - 0{,}300$
$= 0{,}698$

b. $P(\text{Nenhuma}) = 1 - 0{,}698 = 0{,}302$

30. a. $P(A \mid B) = \dfrac{P(A \cap B)}{P(B)} = \dfrac{0{,}40}{0{,}60} = 0{,}6667$

b. $P(B \mid A) = \dfrac{P(A \cap B)}{P(A)} = \dfrac{0{,}40}{0{,}50} = 0{,}80$

c. Não, porque $P(A \mid B) \neq P(A)$

32. a.

	Sim	Não	Total
18 a 34 anos	0,375	0,085	0,46
A partir de 35 anos	0,475	0,065	0,54
Total	0,850	0,150	1,00

b. 46%, 18 a 34 anos; 54% a partir de 35 anos
c. 0,15
d. 0,1848
e. 0,1204
f. 0,5677
g. Maior probabilidade de "Não" para as idades de 18 a 34 anos

33. a.

	Razão para Matricular-se			
	Qualidade	Custo/ Conveniência	Outros	Total
Tempo integral	0,218	0,204	0,039	0,461
Tempo parcial	0,208	0,307	0,024	0,539
Total	0,426	0,511	0,063	1,000

a. A maior probabilidade é que um estudante cite o custo ou conveniência como a primeira razão (probabilidade = 0,511); a qualidade da escola é a razão citada pelo segundo maior número de estudantes (probabilidade = 0,426)
c. $P(\text{qualidade} \mid \text{tempo integral}) = 0{,}218/0{,}461 = 0{,}473$
d. $P(\text{qualidade} \mid \text{tempo parcial}) = 0{,}208/0{,}539 = 0{,}386$
e. Quanto à independência, devemos ter $P(A)P(B) = P(A \cap B)$; da tabela,
$P(A \cap B) = 0{,}218$, $P(A) = 0{,}461$, $P(B) = 0{,}426$
$P(A)P(B) = (0{,}461)(0{,}426) = 0{,}196$
Uma vez que $P(A)P(B) \neq P(A \cap B)$, os eventos não são independentes

34. a. 0,44
b. 0,15
c. 0,136
d. 0,106
e. 0,0225
f. 0,0025

36. a. 0,7921
b. 0,9879
c. 0,0121
d. 0,3364, 0,8236, 0,1764
Não cometer falta em Reggie Miller

38. a. 0,0209
b. 0,0141; 0,027
c. Não
d. 0,0202; 0,0458
e. Sim

39. a. Sim, porque $P(A_1 \cap A_2) = 0$
b. $P(A_1 \cap B) = P(A_1)P(B \mid A_1) = 0{,}40(0{,}20) = 0{,}08$
$P(A_2 \cap B) = P(A_2)P(B \mid A_2) = 0{,}60(0{,}05) = 0{,}03$
c. $P(B) = P(A_1 \cap B) + P(A_2 \cap B) = 0{,}08 + 0{,}03 = 0{,}11$
d. $P(A_1 \mid B) = \dfrac{0{,}08}{0{,}11} = 0{,}7273$
$P(A_2 \mid B) = \dfrac{0{,}03}{0{,}11} = 0{,}2727$

40. a. 0,10; 0,20; 0,09
b. 0,51
c. 0,26; 0,51; 023

42. M = pagamento não efetuado
D_1 = cliente inadimplente
D_2 = cliente não-inadimplente
$P(D_1) = 0{,}05$, $P(D_2) = 0{,}95$, $P(M \mid D_2) = 0{,}2$, $P(M \mid D_1) = 1$

a. $P(D_1 \mid M) = \dfrac{P(D_1)P(M \mid D_1)}{P(D_1)P(M \mid D_1) + P(D_2)P(M \mid D_2)}$
$= \dfrac{(0{,}05)(1)}{(0{,}05)(1) + (0{,}95)(0{,}2)}$
$= \dfrac{0{,}05}{0{,}24} = 0{,}21$

b. Sim, a probabilidade de inadimplência é maior que 20

44. a. 0,47; 0,53; 0,50, 0,45
b. 0,4963
c. 0,4463
d. 47%, 53%

46. a. 0,68
b. 52
c. 10

48. a. 315
b. 0,29
c. Não
d. Republicanos

50. a. 0,76
b. 0,24

52. b. 0, 2022
c. 0,4618
d. 0,4005

54. a. 0,49
b. 0,44
c. 0,54
d. Não
e. Sim

56. a. 0,25
b. 0,125
c. 0,0125

d. 0,10
e. Não

58. 3,44%

60. **a.** 0,40
b. 0,67

Capítulo 5

1. **a.** Cara, Cara, (H, H)
Cara, Coroa (H, T)
Coroa, Cara (T, H)
Coroa, Coroa (T, T)
b. x = número de "caras" ao jogar a moeda duas vezes

c.

Resultado	Valores de x
(H, H)	2
(H, T)	1
(T, H)	1
(T, T)	0

d. Discreta; pode assumir três valores: 0, 1 e 2

2. **a.** x = tempo em minutos para montar o produto
b. Qualquer valor positivo: $x > 0$
c. Contínua

3. Admitamos que S (Sim) = o cargo é oferecido
N (Não) = o cargo não é oferecido
a. $S = \{(S, S, S), (S, S, N), (S, N, S), (S, N, N), (N, S, S), (N, S, N), (N, N, S), (N, N, N)\}$
b. Admitamos que N = número de ofertas feitas; N é uma variável aleatória discreta
c.

Resultado Experimental	(S, S, S)	(S, S, N)	(S, N, S)	(S, N, N)	(N, S, S)	(N, S, S)	(N, N, S)	(N, N, N)
Valor de N	3	2	2	1	2	1	1	0

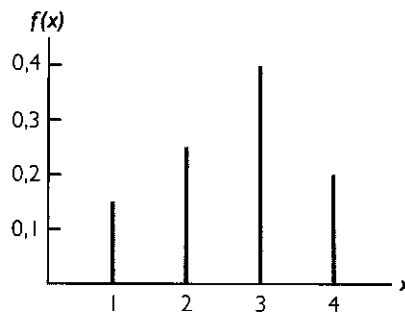
4. $x = 0, 1, 2, \ldots, 12$

6. **a.** 0, 1, 2, . . ., 20; discreta
b. 0, 1, 2, . . .; discreta
c. 0, 1, 2, . . ., 50; discreta
d. $0 \leq x \leq 8$; contínua
e. $x > 0$; contínua

7. **a.** $f(x) \geq 0$ para todos os valores de x
$Sf(x) = 1$; portanto, é uma distribuição de probabilidade válida
b. A probabilidade de $x = 30$ é $f(30) = 0{,}25$
c. A probabilidade de $x \leq 25$ é $f(20) + f(25) = 0{,}20 + 0{,}15 = 0{,}35$
d. A probabilidade de $x > 30$ é $f(35) = 0{,}40$

8. **a.**

x	$f(x)$
1	3/20 = 0,15
2	5/20 = 0,25
3	8/20 = 0,40
4	4/20 = 0,20
Total	1,00

b.

c. $f(x) \geq 0$ para $x = 1, 2, 3, 4$
$\Sigma f(x) = 1$

10. **a.**

x	1	2	3	4	5
$f(x)$	0,05	0,09	0,03	0,42	0,41

b.

x	1	2	3	4	5
$f(x)$	0,04	0,10	0,12	0,46	0,28

c. 0,83
d. 0,28
e. Altos executivos mais satisfeitos

12. **a.** Sim
b. 0,65

14. **a.** 0,05
b. 0,70
c. 0,40

16. **a.**

y	$f(y)$	$yf(y)$
2	0,20	0,40
4	0,30	1,20
7	0,40	2,80
8	0,10	0,80
Totais	1,00	5,20

$E(y) = \mu = 5{,}20$

b.

y	$y - \mu$	$(y - \mu)^2$	$f(y)$	$(y - \mu)^2 f(y)$
2	−3,20	10,24	0,20	2,048
4	−1,20	1,44	0,30	0,432
7	1,80	3,24	0,40	1,296
8	2,80	7,84	0,10	0,784
			Total	4,560

$\text{Var}(y) = 4{,}56$
$\sigma = \sqrt{4{,}56} = 2{,}14$

18. **a/b**

x	$f(x)$	$xf(x)$	$x - \mu$	$(x - \mu)^2$	$(x - \mu)^2 f(x)$
0	0,04	0,00	−1,84	3,39	0,12
1	0,34	0,34	−0,84	0,71	0,24
2	0,41	0,82	0,16	0,02	0,01
3	0,18	0,53	1,16	1,34	0,24
4	0,04	0,15	2,16	4,66	0,17
Total	1,00	1,84			0,79
		↑			↑
		$E(x)$			$Var(x)$

c/d.

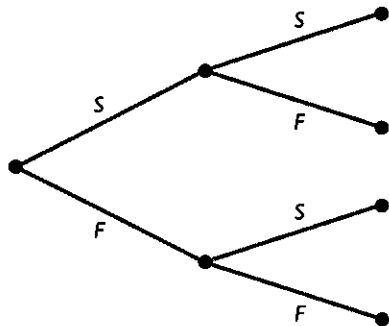
y	$f(y)$	$yf(y)$	$y-\mu$	$(y-\mu)^2$	$y-\mu^2 f(y)$
0	0,00	0,00	−2,93	8,58	0,01
1	0,03	0,03	−1,93	3,72	0,12
2	0,23	0,45	−0,93	0,86	0,20
3	0,52	1,55	0,07	0,01	0,00
4	0,22	0,90	1,07	1,15	0,26
Total	1,00	2,93 ↑ $E(y)$			0,59 ↑ $Var(y)$

e. O número de quartos de dormir nas casas ocupadas pelos proprietários é maior que nas casas ocupadas por inquilinos; o número esperado de quartos de dormir é 1,09 = 2,93 − 1,84 maior, e a variabilidade no número de quartos de dormir é menor no que diz respeito às casas ocupadas pelos proprietários

20. a. 166
b. −94; a preocupação é proteger-se das despesas decorrentes de um grande acidente

22. a. 445
b. Prejuízo de US$ 1.250

24. a. Média: 145; grande: 140
b. Média: 2.725; grande: 12.400

25. a.

b. $f(1) = \binom{2}{1}(0,4)^1(0,6)^1 = \frac{2!}{1!1!}(0,4)(0,6) = 0,48$

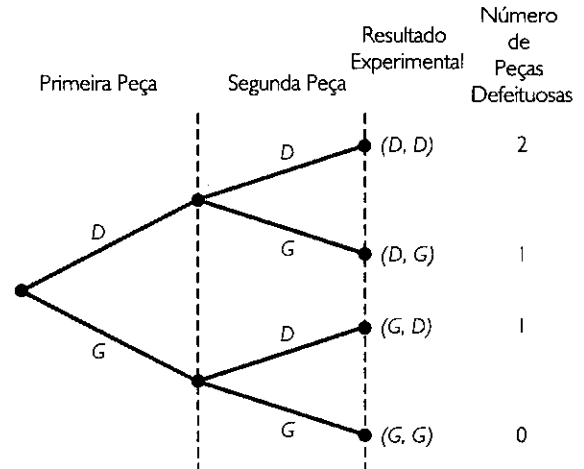
c. $f(0) = \binom{2}{0}(0,4)^0(0,6)^2 = \frac{2!}{0!2!}(1)(0,36) = 0,36$

d. $f(2) = \binom{2}{2}(0,4)^2(0,6)^0 = \frac{2!}{2!0!}(0,16)(0,1) = 0,16$

e. $P(x \geq 1) = f(1) + f(2) = 0,48 + 0,16 = 0,64$
f. $E(x) = np = 2(0,4) = 0,8$
$\text{Var}(x) = np(1-p) = 2(0,4)(0,6) = 0,48$
$\sigma = \sqrt{0,48} = 0,6928$

26. a. $f(0) = 0,3487$
b. $f(2) = 0,1937$
c. 0,9298
d. 0,6513
e. 1
f. $\sigma^2 = 0,9000$, $\sigma = 0,9487$

28. a. 0,2789
b. 0,4181
c. 0,0733

30. a. A probabilidade de uma peça defeituosa ser produzida deve ser igual a 0,03 para cada peça selecionada; as peças devem ser selecionadas independentemente.
b. Admitamos que D = peça defeituosa
G = peça sem defeito

c. Dois resultados representam exatamente um defeito
d. $P(\text{sem defeito}) = (0,97)(0,97) = 0,9409$
$P(1\text{ defeito}) = 2(0,03)(0,97) = 0,582$
$P(2\text{ defeitos}) = (0,03)(0,03) = 0,0009$

32. a. 0.90
b. 0,99
c. 0,999
d. Sim

34. a. 0,0634
b. 0,0634
c. 0,9729

38. a. $f(x) = \frac{3^x e^{-3}}{x!}$
b. 0,2241
c. 0,1494
d. 0,8008

39. a. $f(x) = \frac{2^x e^{-2}}{x!}$
b. μ = seis ocorrências em três períodos
c. $f(x) = \frac{6^x e^{-6}}{x!}$
d. $f(2) = \frac{2^2 e^{-2}}{2!} = \frac{4(0,1353)}{2} = 0,2706$
e. $f(6) = \frac{6^6 e^{-6}}{6!} = 0,1606$
f. $f(5) = \frac{4^5 e^{-4}}{5!} = 0,1563$

40. a. $\mu = 48(5/60) = 4$
$f(3) = \frac{4^3 e^{-4}}{3!} = \frac{(64)(0,0183)}{6} = 0,1952$

b. $\mu = 48(15/50) = 12$

$$f(10) = \frac{12^{10}e^{-12}}{10!} = 0{,}1048$$

c. $\mu = 48(5/60) = 4$; pode-se esperar que haja quatro chamadas telefônicas em espera depois de cinco minutos

$f(0) = \dfrac{4^0e^{-4}}{0!} = 0{,}0183$; a probabilidade de não haver nenhuma espera depois de cinco minutos é 0,0183

d. $m = 48(3/60) = 2{,}4$

$f(0) = \dfrac{2.4^0e^{-2.4}}{0!} = 0{,}0907 = 0{,}0907$; a probabilidade de nenhuma interrupção em três minutos é 0,0907

42. a. $f(0) = \dfrac{7^0e^{-7}}{0!} = e^{-7} = 0{,}0009$

b. probabilidade $= 1 - [f(0) + f(1)]$

$$f(1) = \frac{7^1e^{-7}}{1!} = 7e^{-7} = 0{,}0064$$

probabilidade $= 1 - [0{,}0009 + 0{,}0064] = 0{,}9927$

c. $\mu = 3{,}5$

$$f(0) = \frac{3.5^0e^{-3.5}}{0!} = e^{-3.5} = 0{,}0302$$

probabilidade $= 1 - f(0) - 0{,}0302 = 0{,}9698$

d. probabilidade $= 1 - [f(0) + f(1) + f(2) + f(3) + f(4)]$
$= 1 - [0{,}0009 + 0{,}0064 + 0{,}0223 + 0{,}0521 + 0{,}0912]$
$= 0{,}8271$

44. a. $\mu = 1{,}25$
b. 0,2865
c. 0,3581
d. 0,3554

46. a. $f(1) = \dfrac{\binom{3}{1}\binom{10-3}{4-1}}{\binom{10}{4}} = \dfrac{\left(\frac{3!}{1!2!}\right)\left(\frac{7!}{3!4!}\right)}{\frac{10!}{4!6!}}$

$$= \frac{(3)(35)}{210} = 0{,}50$$

b. $f(2) = \dfrac{\binom{3}{2}\binom{10-3}{2-2}}{\binom{10}{2}} = \dfrac{(3)(1)}{45} = 0{,}067$

c. $f(0) = \dfrac{\binom{3}{0}\binom{10-3}{2-0}}{\binom{10}{2}} = \dfrac{(1)(21)}{45} = 0{,}4667$

d. $f(2) = \dfrac{\binom{3}{2}\binom{10-3}{4-2}}{\binom{10}{4}} = \dfrac{(3)(21)}{210} = 0{,}30$

48. a. 0,5250
b. 0,1833

50. $N = 60$, $n = 10$

a. $r = 20$, $x = 0$

$$f(0) = \frac{\binom{20}{0}\binom{40}{10}}{\binom{60}{10}} = \frac{(1)\left(\frac{40!}{10!30!}\right)}{\frac{60!}{10!50!}}$$

$$= \left(\frac{40!}{10!30!}\right)\left(\frac{10!50!}{60!}\right)$$

$$= \frac{40\cdot39\cdot38\cdot37\cdot36\cdot35\cdot34\cdot33\cdot32\cdot31}{60\cdot59\cdot58\cdot57\cdot56\cdot55\cdot54\cdot53\cdot52\cdot51}$$

$$\approx .01$$

b. $r = 20$, $x = 1$

$$f(1) = \frac{\binom{20}{1}\binom{40}{9}}{\binom{60}{10}} = 20\left(\frac{40!}{9!31!}\right)\left(\frac{10!50!}{60!}\right)$$

$$\approx .07$$

c. $1 - f(0) - f(1) = 1 - 0{,}08 = 0{,}92$

d. A mesma probabilidade de um dos empregados ser da fábrica do Havaí; no item (b), era de aproximadamente 0,07

52. a. 0,5333
b. 0,6667
c. 0,7778
d. $n = 7$

54. a.

x	1	2	3	4	5
$f(x)$	0,24	0,21	0,10	0,21	0,24

b. 3,00; 2,34
c. Debêntures: $E(x) = 1{,}36$, $\text{Var}(x) = 0{,}23$
Fundos de ações: $E(x) = 4$, $\text{Var}(x) = 1$

56. a. 0,0596
b. 0,3585
c. 100
d. 9,7468

58. a. 0,9510
b. 0,0480
c. 0,0490

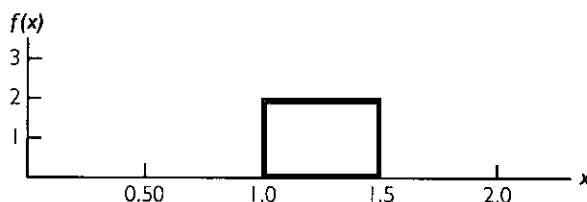
60.
a. 240
b. 12,9615
c. 12,9615

62. 0,1912

64. a. 0,2240
b. 0,5767

66. a. 0,4667
b. 0,4667
c. 0,0667

Capítulo 6

1. a.

b. $P(x = 1{,}25) = 0$; a probabilidade de qualquer ponto em particular é zero porque a área sob a curva acima de qualquer ponto em particular é zero.
c. $P(1{,}0 \leq x \leq 1{,}25) = 2(0{,}25) = 0{,}50$
d. $P(1{,}20 < x < 1{,}5) = 2(0{,}30) = 0{,}60$

2. **b.** 0,50
c. 0,60
d. 15
e. 8,33

4. **a.**

b. $P(0{,}25 < x < 0{,}75) = 1(0{,}50) = 0{,}50$
c. $P(x \leq 0{,}30) = 1(0{,}30) = 0{,}30$
d. $P(x > 0{,}60) = 1(0{,}40) = 0{,}40$

6. **a.** 0,40
b. 0,64
c. 0,68

10. **a.** 0,3413
b. 0,4332
c. 0,4772
d. 0,4938

12. **a.** 0,2967
b. 0,4418
c. 0,3300
d. 0,5910
e. 0,8849
f. 0,2389

13. **a.** 0,6879 – 0,0239 = 0,6640
b. 0,8888 – 0,6985 = 0,1903
c. 0,9599 – 0,8508 = 0,1091

14. **a.** $z = 1{,}96$
b. $z = 0{,}61$
c. $z = 1{,}12$
d. $z = 0{,}44$

15. **a.** Procure na tabela uma área igual a 0,5000 – 0,2119 = 0, 2881; $z = 0{,}80$ destaca uma área igual a 0,2119 na cauda superior; desse modo, para uma área igual a 0,2119 na cauda inferior, $z = -0{,}80$
b. Procure na tabela uma área igual a 0,9030/2 = 0,4515; $z =$ 1,66
c. Procure na tabela uma área igual a 0,2052/2 = 0,1026; z =0,26
d. Procure na tabela uma área igual a 0,4948; $z = 2{,}56$
e. Procure na tabela uma área igual a 0,1915; uma vez que o valor que procuramos está abaixo da média, o valor *z negativo; assim*, $z = -0{,}50$

16. **a.** $z = 2{,}33$
b. $z = 1{,}96$
c. $z = 1{,}645$
d. $z = 1{,}28$

18. $\mu = 30$ e $s = 8{,}2$
a. Para $x = 40$, $z = \dfrac{40 - 30}{8{,}2} = 1{,}22$
$P(z \leq 1{,}22) = 0{,}5000 + 0{,}3888 = 0{,}8888$
$P(x \geq 40) = 1{,}000 - 0{,}8888 = 0{,}1112$
b. Para $x = 20$, $z = \dfrac{20 - 30}{8{,}2} = -1{,}22$
$P(z > -1{,}22) = 0{,}5000 + 0{,}3888 = 0{,}8888$
$P(x \leq 20) = 1{,}000 - 0{,}8888 = 0{,}1112$
c. Um valor z igual a 1,28 destaca uma área de aproximadamente 10% na cauda superior
$x = 30 + 8{,}2(1{,}28)$
$= 40{,}50$
Um preço de US$ 40,50 ou mais por ação colocará a empresa entre as 10% maiores.

20. **a.** 0,0885
b. 12,51%
c. 93,8 horas ou mais

22. **a.** 0,4194
b. US$ 517,44 ou mais
c. 0,0166

24. **a.** 902,75; 114,185
b. 0,1841
c. 0,1977
d. 1.091 milhão

26. **a.** $\mu = np = 100(0{,}20) = 20$
$\sigma^2 = np(1 - p) = 100(0{,}20)(0{,}80) = 16$
$\sigma = \sqrt{16} = 4$
b. Sim, porque $np = 20$ e $b(1 - p) = 80$
c. $P(23{,}5 \leq x \leq 24{,}5)$
$z = \dfrac{24{,}5 - 20}{4} = +1{,}13$ Área = 0,3708
$z = \dfrac{23{,}5 - 20}{4} = +0{,}88$ Área = 0,3106
$P(23{,}5 \leq x \leq 24{,}5) = 0{,}3708 - 0{,}3106 = 0{,}0602$
d. $P(17{,}5 \leq x \leq 22{,}5)$
$z = \dfrac{17{,}5 - 20}{4} = -0{,}63$ Área = 0,2357
$z = \dfrac{22{,}5 - 20}{4} = +0{,}63$ Área = 0,2357
$P(17{,}5 = x = 22{,}5) = 0{,}2357 + 0{,}2357 = 0{,}4714$
e. $P(x = 15{,}5)$
$z = \dfrac{15{,}5 - 20}{4} = -1{,}13$ Área = 0,3708
$P(x = 15{,}5) = 0{,}5000 - 0{,}3708 = 0{,}1292$

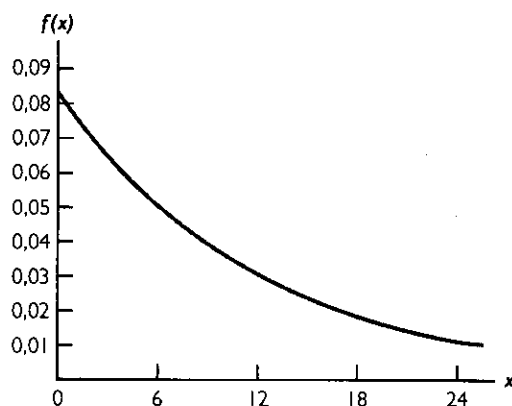
28. **a.** 0,1867
b. 125
c. É um lance de sorte

30. **a.** 220
b. 0,0392
c. 0,8962

32. **a.** 0,5276
b. 0,3935
c. 0,4724
d. 0,1341

33. **a.** $P(x = x_0) = 1 - e^{-x_0/3}$
b. $P(x = 2) = 1 - e^{-2/3} = 1 - 0{,}5134 = 0{,}4866$
c. $P(x \geq 3) = 1 - P(x = 3) = 1 - (1 - e^{-3/3})$
$= e^{-1} = 0{,}3679$
d. $P(x = 5) = e^{-5/3} = 1 - 0{,}1889 = 8111$
e. $P(2 = x = 5) = P(x = 5) - p(x = 2)$
$= 0{,}8111 - 0{,}4866 = 0{,}3245$

34. **a.** 0,3935
b. 0,2231
c. 0,3834

35. **a.**

b. $P(x \leq 12) = 1 - e^{-12/12} = 1 - 0{,}3679 = 0{,}6321$
c. $P(x \leq 6) = 1 - e^{-6/12} = 1 - 0{,}6065 = 0{,}3935$
d. $P(x \geq 30) = 1 - P(x < 30)$
$= 1 - (1 - e^{-30/12})$
$= 0{,}0821$

36. **a.** 50 horas
b. 0,3935
c. 0,1353

38. **a.** $f(x) = 30e^{-30x}$
b. 0,0821
c. 0,7135

40. **a.** US$ 3.780 ou menos
b. 19,22%
c. US$ 8.167,50

42. **a.** 3.229
b. 0,2244
c. US$ 12.382 ou mais

44. **a.** 0,0228
b. US$ 50,00

46. **a.** 38,3%
b. 3,59%, na melhor das hipóteses; 96,41%, na pior.
c. 38,21%

48. $\mu = 19{,}23$ onças (568,70 mL)

50. **a.** Perder US$ 240,00
b. 0,1788
c. 0,3557
d. 0,0594

52. **a.** $1/7$ de minuto
b. $7e^{-7x}$
c. 0,0009
d. 0,2466

54. **a.** 2 minutos
b. 0,2212
c. 0,3935
d. 0,0821

Capítulo 7

1. **a.** AB, AC, AD, AE, BC, BD, BE, CD, CE, DE
b. Com dez amostras, cada uma tem uma probabilidade de $1/10$
c. E e C, porque 8 e 0 não se aplicam; 5 identifica E; 7 não se aplica; pula-se 5 porque E já está na amostra; 3 identifica C; 2 não é necessário porque o tamanho 2 da amostra já está completo

2. 22, 147, 229, 289

3. 459, 147, 385, 113, 340, 401, 215, 2, 33, 348

4. **a.** Nasdaq 100, Oracle, Microsoft, Lucent, Applied Materials
b. 252

6. 2.782, 493, 825, 1.807, 289

8. Maryland, Iowa, Estado da Flórida, Virgínia, Pittsburgh, Oklahoma.

10. **a.** finita
b. infinita
c. infinita
d. infinita
e. finita

11. **a.** $\bar{x} = \dfrac{\Sigma x_i}{n} = \dfrac{54}{6} = 9$
b. $s = \sqrt{\dfrac{\Sigma(x_i - \bar{x})^2}{n - 1}}$

$\Sigma(x_i - x)^2 = (-4)^2 + (-1)^2 + 1^2 + (-2)^2 + 1^2 + 5^2$
$= 48$

$s = \sqrt{\dfrac{48}{6 - 1}} = 3{,}1$

12. **a.** 0,50
b. 0,3667

13. **a.** $\bar{x} = \dfrac{\Sigma x_i}{n} = \dfrac{465}{5} = 93$

b.

	x_i	$(x_i - x)$	$(x_i - x)^2$
	94	+1	1
	100	+7	49
	85	−8	64
	94	+1	1
	92	−1	1
Totais	465	0	116

$$s = \sqrt{\frac{\Sigma(x_i - \bar{x})^2}{n-1}} = \sqrt{\frac{116}{4}} = 5{,}39$$

14. **a.** 0,45
b. 0,15
c. 0,45

16. **a.** 0,10
b. 20
c. 0,72

18. **a.** 200
b. 5
c. Normal, com $E((\bar{x}))= 200$ e $\sigma_{\bar{x}} = 5$
d. A distribuição de probabilidade de $\bar{x}$

19. **a.** A distribuição amostral é normal com

$E(\bar{x}) = \mu = 200$

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{50}{\sqrt{100}} = 5$$

Para +5, $(\bar{x} - \mu) = 5$,

$$z = \frac{\bar{x} - \mu}{\sigma_{\bar{x}}} = \frac{5}{5} = 1$$

Área = 2(0,3413) = 0,6826

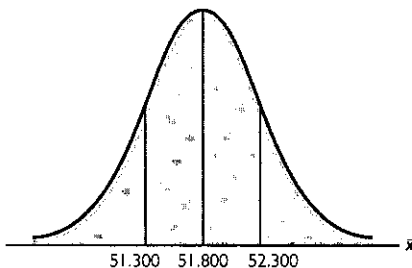
b. Para ±10, $(\bar{x} - \mu) = 10$,

$$z = \frac{\bar{x} - \mu}{\sigma_{\bar{x}}} = \frac{10}{5} = 2$$

Área = 2(0,4772) = 0,9544

20. 3,54; 2,50; 2,04; 1,77
$\sigma_{\bar{x}}$ decresce à medida que n se eleva

22. **a.** Normal, com $E(\bar{x}) = 51.800$ e $\sigma_{\bar{x}} = 516{,}40$
b. $\sigma_{\bar{x}}$ decresce para 365,15
c. $\sigma_{\bar{x}}$ decresce à medida que n se eleva

23. **a.**

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{4.000}{\sqrt{60}} = 516{,}40$$

$$z = \frac{52.300 - 51.800}{516{,}40} = +0{,}97$$

Área = 2(0,3340) = 0,6680

b. $$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{4.000}{\sqrt{120}} = 365{,}15$$

$$z = \frac{52.300 - 51.800}{365{,}15} = +1{,}37$$

Área = 2(0.4147) = 0.8294

24. **a.** Normal, com $E(\bar{x}) = 4.260$ e $\sigma_{\bar{x}} = 127{,}28$
b. 0,95
c. 0,5704

26. **a.** 0,5034; 0,6212; 0,7888; 0,9232; 0,9876
b. Maior probabilidade dentro de ±250

28. **a.** Normal, com $E(\bar{x}) = 687$ e $\sigma_{\bar{x}} = 34{,}29$
b. 0,9964
c. 0,5346
d. Aumentar o tamanho da amostra

30. **a.** $n/N = 0{,}01$; não
b. 1,29; 1,30; pouca diferença
c. 0,8764

32. **a.** $E(\bar{p}) = 0{,}40$

$$\sigma_{\bar{p}} = \sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{(0{,}40)(0{,}60)}{200}} = 0{,}0346$$

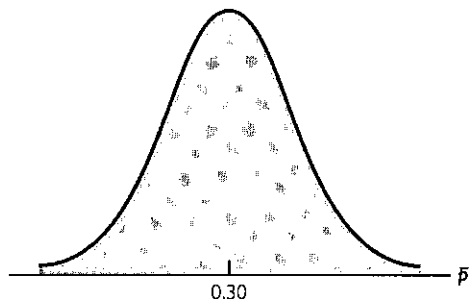
$$z = \frac{\bar{p} - p}{\sigma_{\bar{p}}} = \frac{0{,}03}{0{,}0346} = 0{,}87$$

Área = 2(0,3078) = 0,6156

b. $$\frac{\bar{p} - p}{\sigma_{\bar{p}}} = \frac{0{,}05}{0{,}0346} = 1{,}44$$

Área = 2 (0,4251) = 0,8502

34. **a.** 0,6156
b. 0,7814
c. 0,9488
d. 0,9942
e. Maior probabilidade com n maior

35. **a.** $$\sigma_{\bar{p}} = \sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{0{,}30(0{,}70)}{100}} = 0{,}0458$$

A distribuição normal é apropriada porque $np = 100(0{,}30) = 30$ e $n(1-p) = 100(0{,}70) = 70$ são ambos maiores que 5

b. $P(0{,}20 \leq \bar{p} \leq 0{,}40) = ?$

$$z = \frac{0{,}40 - 0{,}30}{0{,}0458} = 2{,}18$$

Área = 2(0,4854) = 0,9708

c. $P(0{,}25 \leq \bar{p} \leq 0{,}35) = ?$

$$z = \frac{0{,}35 - 0{,}30}{0{,}0458} = 1{,}09$$

Área = 2(0,3621) = 0,7242

36. **a.** Normal, com $E(\bar{p}) = 0,56$ e $\sigma_{\bar{p}} = 0,287$
b. 0,7062
c. 0,8612; 0,9438

38. **a.** Normal, com $E(\bar{p}) = 0,56$ e $\sigma_{\bar{p}} = 0,0248$
b. 0,5820
c. 0,8926

40. **a.** Normal, com $E(\bar{p}) = 0,76$ e $\sigma_{\bar{p}} = 0,0214$
b. 0, 8384
c. 0,9452

42. *112, 145*, 73, 324, 293, 875, 318, 618

44. **a.** Normal, com $E(\bar{x}) = 115,50$ e $\sigma_{\bar{x}} = 5,53$
b. 0,9298
c. 0,0026

46. **a.** 707
b. 0,50
c. 0,8414
d. 0,9544

48. **a.** 625
b. 0,7888

50. **a.** Normal, com $E(\bar{p}) = 0,305$ e $\sigma_{\bar{p}} = 0,0326$
b. 0,7814
c. 0,4582

52. **a.** 0,9606
b. 0,0495

54. **a.** 48
b. Normal, com $E(\bar{p}) = 0,25$ e $\sigma_{\bar{p}} = 0,0625$
c. 0,2119

Capítulo 8

2. Use $\bar{x} \pm z_{\alpha/2}(\sigma/\sqrt{n})$
a. $32 \pm 1,645\ (6/\sqrt{50})$
$32 \pm 1,4$; 30,6 a 33,4
b. $32 \pm 1,96(6/\sqrt{50})$
$32 \pm 1,66$; 30,34 a 33,66
c. $32 \pm 2,576(6/\sqrt{50})$
$32 \pm 2,19$; 29,81 a 34,19

4. 54

5. **a.** $1,96\sigma/\sqrt{n} = 1,96(5/\sqrt{49}) = 1,40$
b. $24,80 \pm 1,40$; 23,40 a 26,20

6. 8,1 a 8,9

8. **a.** A população é, no mínimo, aproximadamente normal
b. 3,1
c. 4,1

10. **a.** US$ 113.638 a US$ 124.672
b. US$ 112.581 a US$ 125.729
c. US$ 110.515 a US$ 127.795
d. A amplitude aumenta quando o nível de confiança sobe

12. **a.** 2,179
b. –1,676
c. 2,457
d. –1,708 e 1,708
e. –2,014 e 2,014

13. **a.** $\bar{x} = \frac{\Sigma x_i}{n} = \frac{80}{8} = 10$
b. $s = \sqrt{\frac{\Sigma(x_i - \bar{x})^2}{n-1}} = \sqrt{\frac{84}{7}} = 3,46$
c. $t_{0,025}\left(\frac{s}{\sqrt{n}}\right) = 2,365\left(\frac{3,46}{\sqrt{8}}\right) = 2,9$
d. $\bar{x} \pm t_{0,025}\left(\frac{s}{\sqrt{n}}\right)$
$10 \pm 2,9$ (7,1 a 12,9)

14. **a.** 21,5 a 23,5
b. 21,3 a 23,7
c. 20,9 a 24,1
d. Uma margem de erro maior e um intervalo mais amplo

15. $\bar{x} \pm t_{\alpha/2}(s/\sqrt{n})$
confiança de 90%: $gl = 64$ e $t_{0,05} = 1,669$
$19,5 \pm 1,669\left(\frac{5,2}{\sqrt{65}}\right)$
$19,5 \pm 1,08$ (18,42 a 20,58)
confiança de 95%: $gl = 64$ e $t_{0,025} = 1,998$
$19,5 \pm 1,998\left(\frac{5,2}{\sqrt{65}}\right)$
$19,5 \pm 1,29$ (18,21 a 20,79)

16. **a.** 1,69
b. 47,31 a 50,69
c. Menos horas e custos mais elevados para a United Airlines

18. **a.** 3,8
b. 0,84
c. 2,96 a 4,64
d. Maior n na próxima vez

20. 6,28 a 6,78

22. **a.** 3,35
b. 2,40 a 4,30

24. **a.** Valor planejado de $\sigma = \frac{\text{Intervalo}}{4} = \frac{36}{4} = 9$
b. $n = \frac{z^2_{0,025}\sigma^2}{E^2} = \frac{(1,96)^2(9)^2}{(3)^2} = 34,57$; use $n = 35$
c. $n = \frac{(1,96)^2(9)^2}{(2)^2} = 77,79 = 77,79$; use $n = 78$

25. **a.** Use $n = \frac{z^2_{\alpha/2}\sigma^2}{E^2}$
$n = \frac{(1,96)^2(6,82)^2}{(1,5)^2} = 79,41$; use $n = 80$
b. $n = \frac{(1,645)^2(6,82)^2}{(2)^2} = 31,47$; use $n = 32$

26. **a.** 340
b. 1.358
c. 8.887

28. **a.** 343
b. 487
c. 840
d. n torna-se maior; com 99% de confiança, não recomendaria.

30. 81

31. **a.** $\bar{p} = \frac{100}{400} = 0{,}25$

b. $\sqrt{\frac{\bar{p}(1-\bar{p})}{n}} = \sqrt{\frac{0{,}25(0{,}75)}{400}} = 0{,}0217$

c. $\bar{p} \pm z_{0,025}\sqrt{\frac{\bar{p}(1-\bar{p})}{n}}$
$0{,}25 \pm 1{,}96(0{,}0217)$
$0{,}25 \pm 0{,}0424$; 0,2076 a 0,2924

32. **a.** 0,6733 a 0,7267
b. 0,6682 a 0,7318

34. 1.068

35. **a.** $\bar{p} = \frac{281}{611} = .4599$ (46%)

b. $z_{0,05}\sqrt{\frac{\bar{p}(1-\bar{p})}{n}} = 1{,}645\sqrt{\frac{4599(1-0{,}4599)}{611}} = 0{,}0332$

c. $\bar{p} \pm 0{,}0332$
$0{,}4599 \pm 0{,}0332$ (0,4267 a 0,4931)

36. **a.** 0,4393
b. 0,3870 a 0,4916

38. **a.** 0,0430
b. 0,2170 a 0,3030
c. 822

39. **a.** $n = \frac{1{,}96^2 p^*(1-p^*)}{E^2}$

$n = \frac{1{,}96^2(0{,}33)(0{,}67)}{(0{,}03)^2} = 943{,}75$; use $n = 944$

b. $n = \frac{2{,}576^2(0{,}33)(0{,}67)}{(0{,}03)^2} = 1.630{,}19$; use $n = 1.631$

40. 0,0267; (0,8333 a 0,8867)

42. **a.** 0,0442
b. 601, 1.068, 2.401, 9.604

44. **a.** 2.009
b. 47.991 a 52.009

46. **a.** 998
b. US$ 24.479 a US$ 26.455
c. US$ 24.479 a US$ 26.455
c. US$ 93,5 milhões
d. Sim, US$ 21,4 (30%) a mais de *O Mundo Perdido – Jurassic Park*

48. **a.** 14 minutos
b. 13,38 a 14,62
c. 32 por dia
d. Redução do quadro de funcionários

50. 37

52. 176

54. **a.** 0,5420
b. 0,0508
c. 0,4912 a 0,5928

56. **a.** 0,68
b. 0,6391 a 0,7209

58. **a.** 1.267
b. 1.509

60. **a.** 0,3101
b. 0,2898 a 0,3304
c. 8.219; não, esse tamanho de amostra é desnecessariamente grande

Capítulo 9

2. **a.** H_0: $\mu = 14$
H_a: $\mu > 14$
b. Não há evidências de que o novo plano aumentará as vendas
c. A hipótese de pesquisa $\mu > 14$ é sustentável; o novo plano aumenta as vendas

4. **a.** H_0: $\mu \geq 220$
H_a: $\mu < 220$

5. **a.** Rejeitar H_0: $\mu = 56{,}2$ quando ela é verdadeira
b. Aceitar H_0: $\mu = 56{,}2$ quando ela é falsa

6. **a.** H_0: $\mu = 1$
H_a: $\mu > 1$
b. Afirmar que $\mu > 1$ quando isso não é verdadeiro
c. Afirmar que $\mu = 1$ quando isso não é verdadeiro

8. **a.** H_0: $\mu \geq 220$
H_a: $\mu < 220$
b. Afirmar que $\mu < 220$ quando isso não é verdadeiro
c. Afirmar que $\mu \geq 220$ quando isso não é verdadeiro

10. **a.** $z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} = \frac{26.4 - 25}{6/\sqrt{40}} = 1{,}48$
b. Área = 0,4306
Valor $p = 0{,}5000 - 0{,}4306 = 0{,}0694$
c. Valor $p > 0{,}01$, não rejeitar H_0
d. Rejeitar H_0 se $z \geq 2{,}33$
$1{,}48 < 2{,}33$, não rejeitar H_0

11. **a.** $z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} = \frac{14{,}15 - 15}{3/\sqrt{50}} = -2{,}00$
b. Área = 0,4772
Valor $p = 2(0{,}5000 - 0{,}4772) = 0{,}0456$
c. Valor $p = 0{,}05$, rejeitar H_0
d. Rejeitar H_0 se $z = -1{,}96$ ou $z \geq 1{,}96$
$-2{,}00 = -1{,}96$, rejeitar H_0

12. **a.** 0,1056; não rejeitar H_0
b. 0,0062; rejeitar H_0
c. ≈ 0; rejeitar H_0
d. 0,7967; não rejeitar H_0

14. **a.** 0,3844; não rejeitar H_0

b. 0,0074; rejeitar H_0
c. 0,0836; não rejeitar H_0

15. a. H_0: $\mu \geq 1.056$
H_a: $\mu < 1.056$
b. $z = \dfrac{\bar{x} - \mu_0}{\sqrt{\sigma/n}} = \dfrac{910 - 1.056}{1.600/\sqrt{400}} = -1,83$
Valor $p = 0,5000 - 0,4664 = 0,0336$
c. Valor $p = 0,05$, rejeitar H_0. A média de restituição do IR para quem faz declarações "de última hora" é inferior a US$ 1.056
d. Rejeitar H_0 se $z = -1,645$
$-1,83 = -1,645$, rejeitar H_0

16. a. H_0: $\mu = 895$
H_a: $\mu > 895$
b. 0,1170
c. Não rejeitar H_0
d. Manter o julgamento; coletar mais dados

18. a. H_0: $\mu = 4,1$
H_a: $\mu \neq 4,1$
b. –2,21; 0,0272
c. Rejeitar H_0

20. a. H_0: $\mu \geq 181.900$
H_a: $\mu < 181.900$
b. –2,93
c. 0,0017
d. Rejeitar H_0

22. a. H_0: $\mu = 8$
H_a: $\mu \neq 8$
b. 0,1706
c. Não rejeitar H_0
d. 7,83 a 8,97; Sim

24. a. $t = \dfrac{\bar{x} - \mu_0}{s/\sqrt{n}} = \dfrac{17 - 18}{4,5/\sqrt{48}} = -1,54$
b. Graus de liberdade $= n - 1 = 47$
A área na cauda inferior se encontra entre 0,05 e 0,10
O valor p (bicaudal) se encontra entre 0,10 e 0,20
c. Valor $p > 0,05$; não rejeitar H_0
d. Com $gl = 47$, $t_{0,025} = 2,012$
Rejeitar H_0 se $t = -2,012$ ou $t \geq 2,012$
$t = -1,54$; não rejeitar H_0

26. a. Entre 0,02 e 0,05; rejeitar H_0
b. Entre 0,01 e 0,02; rejeitar H_0
c. Entre 0,10 e 0,20; não rejeitar H_0

27. a. H_0: $\mu \geq 238$
H_a: $\mu < 238$
b. $t = \dfrac{\bar{x} - \mu_0}{s/\sqrt{n}} = \dfrac{231 - 238}{80/\sqrt{100}} = -1,54$
Graus de liberdade $= n - 1 = 99$
Valor p entre 0,10 e 0,20
c. Valor $p > 0,05$; não rejeitar H_0
Não se pode concluir que a média dos benefícios semanais em Virgínia é menor que a média nacional
d. $gl = 99 t_{0,05} = -1,66$
Rejeitar H_0 se $t = -1,66$
$-0,88 > -1,66$; não rejeitar H_0

28. a. H_0: $\mu = 3.530$
H_a: $\mu > 3.530$
b. Entre 0,005 e 0,01
c. Rejeitar H_0

30. a. H_0: $\mu = 600$
H_a: $\mu \neq 600$
b. Entre 0,20 e 0,40
c. Não rejeitar H_0
d. Um tamanho de amostra maior

32. a. H_0: $\mu = 10.192$
H_a: $\mu \neq 10.192$
b. Entre 0,02 e 0,05
c. Rejeitar H_0

34. a. H_0: $\mu = 2$
H_a: $\mu \neq 2$
b. 2,2
c. 0,52
d. Entre 0,20 e 0,40
e. Não rejeitar H_0

36. a. $z = \dfrac{\bar{p} - p_0}{\sqrt{\dfrac{p_0(1 - p_0)}{n}}} = \dfrac{0,68 - 0,75}{\sqrt{\dfrac{0,75(1 - 0,75)}{300}}} = -2,80$

Valor $p = 0,5000 - 0,4974 = 0,0026$
Valor $p = 0,05$; rejeitar H_0
b. $z = \dfrac{0,72 - 0,75}{\sqrt{\dfrac{0,75(1 - 0,75)}{300}}} = -1,20$

Valor $p = 0,5000 - 0,3849 = 0,1151$
Valor $p > 0,05$; não rejeitar H_0
c. $z = \dfrac{0,70 - 0,75}{\sqrt{\dfrac{0,75(1 - 0,75)}{300}}} = -2,00$

Valor $p = 0,5000 - 0,4772 = 0,0228$
Valor $p = 0,05$; rejeitar H_0
d. $z = \dfrac{0,77 - 0,75}{\sqrt{\dfrac{0,75(1 - 0,75)}{300}}} = 0,80$

Valor $p = 0,5000 + 0,2881 = 0,7881$
Valor $p > 0,05$; não rejeitar H_0

38. a. H_0: $p = 0,64$
H_a: $p \neq 0,64$
b. $\bar{p} = 52/100 = 0,52$
$z = \dfrac{\bar{p} - p_0}{\sqrt{\dfrac{p_0(1 - p_0)}{n}}} = \dfrac{0,52 - 0,64}{\sqrt{\dfrac{0,64(1 - 0,64)}{100}}} = -2,50$

Área = 0,4938
Valor $p = 2(0,5000 - 0,4938) = 0,0124$
c. Valor $p = 0,05$; rejeitar H_0
A proporção difere do valor 0,64 relatado

d. Sim, porque $\bar{p} = 0,52$ indica que menos pessoas acreditam que a marca de supermercado é tão boa quanto a marca conhecida nacionalmente.

40. **a.** 0,2702
b. H_0: $p = 0,22$
H_a: $p > 0,22$
Valor $p \approx 0$; rejeitar H_0
c. Porque ajudam a avaliar a eficácia dos comerciais

42. H_0: $p = 0,24$
H_a: $p > 0,24$
Valor $p = 0,0023$; rejeitar H_0

44. **a.** H_0: $p = 0,51$
H_a: $p \neq 0,51$
b. $\bar{p} = 0,58$, valor $p = 0,0026$
c. Rejeitar H_0

46. **a.** H_0: $\mu = 16$
H_a: $\mu \neq 16$
b. 0,0286; rejeitar H_0
Reajustar a linha de produção
c. 0,2186; não rejeitar H_0
Continuar a operação
d. $z = 2,19$; rejeitar H_0
$z = -1,23$; não rejeitar H_0
Sim, a mesma conclusão

48. **a.** H_0: $\mu = 45.250$
H_a: $\mu \neq 45.250$
b. 0,0034
c. Rejeitar H_0

50. $t = -0,93$
Valor p entre 0,20 e 0,40
Não rejeitar H_0

52. $t = 2,26$
Valor p entre 0,01 e 0,025
Rejeitar H_0

54. **a.** H_0: $p = 0,50$
H_a: $p > 0,50$
b. 0,64
c. 0,0026; rejeitar H_0

56. **a.** H_0: $p = 0,50$
H_a: $p > 0,50$
b. 0,6381
c. 0,0023; rejeitar H_0

58. H_0: $p \geq 0.90$
H_a: $p < 0,90$
Valor $p = 0,0808$
Não rejeitar H_0

Capítulo 10

1. **a.** $\bar{x}_1 - \bar{x}_2 = 13,6 - 11,6 = 2$
b. $z_{\alpha/2} = z_{0,05} = 1,645$

$$\bar{x}_1 - \bar{x}_2 \pm 1,645\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

$$2 \pm 1,645\sqrt{\frac{(2,2)^2}{50} + \frac{(3)^2}{35}}$$

$2 \pm 0,98$ (1,02 a 2,98)

c. $z_{\alpha/2} = z_{0,05} = 1,96$

$$2 \pm 1,96\sqrt{\frac{(2,2)^2}{50} + \frac{(3)^2}{35}}$$

2. **a.** $$z = \frac{(\bar{x}_1 - \bar{x}_2) - D_0}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} = \frac{(25,2 - 22,8) - 0}{\sqrt{\frac{(5,2)^2}{40} + \frac{(6)^2}{50}}} = 2,03$$

b. Valor $p = 0,5000 - 0,4788 = 0,212$
c. Valor $p = 0,05$; rejeitar H_0

4. **a.** $\bar{x}_1 - \bar{x}_2 = 2,04 - 1,72 = 0,32$
b. $$z_{0,025}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} = 1,96\sqrt{\frac{(0,10)^2}{40} + \frac{(0,08)^2}{35}} = 0,04$$
c. $0,32 \pm 0,04$ (0,28 a 0,36)

6. Valor $p = 0,015$
Rejeitar H_0; um aumento

8. **a.** 1,08
b. 0,2802
c. Não rejeitar H_0; não se pode concluir que exista uma diferença.

9. **a.** $\bar{x}_1 - \bar{x}_2 = 22,5 - 20,1 = 2,4$
b.
$$gl = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{1}{n_1 - 1}\left(\frac{s_1^2}{n_1}\right)^2 + \frac{1}{n_2 - 1}\left(\frac{s_2^2}{n_2}\right)^2}$$

$$= \frac{\left(\frac{2,5^2}{20} + \frac{4,8^2}{30}\right)^2}{\frac{1}{19}\left(\frac{2,5^2}{20}\right)^2 + \frac{1}{29}\left(\frac{4,8^2}{30}\right)^2} = 45,8$$

c. $gl = 45$, $t_{0,025} = 2,014$

$$t_{0,025}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} = 2,014\sqrt{\frac{2,5^2}{20} + \frac{4,8^2}{30}} = 2,1$$

d. $2,4 \pm 2,1$ (0,3 a 4,5)

10. **a.** $$t = \frac{(\bar{x}_1 - \bar{x}_2) - 0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{(13.6 - 10.1) - 0}{\sqrt{\frac{5.2^2}{35} + \frac{8.5^2}{40}}} = 2.18$$

b. $$gl = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{1}{n_1 - 1}\left(\frac{s_1^2}{n_1}\right)^2 + \frac{1}{n_2 - 1}\left(\frac{s_2^2}{n_2}\right)^2}$$

$$= \frac{\left(\frac{5,2^2}{35} + \frac{8,5^2}{40}\right)^2}{\frac{1}{34}\left(\frac{5,2^2}{35}\right)^2 + \frac{1}{39}\left(\frac{8,5^2}{40}\right)^2} = 65,7$$

Use $gl = 65$

c. $gl = 65$, a área na cauda superior se encontra entre 0,01 e 0,025
O valor p bicaudal está entre 0,02 e 0,05
d. Valor $p = 0{,}05$; rejeitar H_0

12. a. $\bar{x}_1 - \bar{x}_2 = 22{,}5 - 18{,}6 = 3{,}9$ milhas

b. $$gl = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{1}{n_1 - 1}\left(\frac{s_1^2}{n_1}\right)^2 + \frac{1}{n_2 - 1}\left(\frac{s_2^2}{n_2}\right)^2}$$

$$= \frac{\left(\frac{8{,}4^2}{50} + \frac{7{,}4^2}{40}\right)^2}{\frac{1}{49}\left(\frac{8{,}4^2}{50}\right)^2 + \frac{1}{39}\left(\frac{7{,}4^2}{40}\right)^2} = 87{,}1$$

Use $gl = 87$, $t_{0{,}025} = 1{,}988$

$$3{,}9 \pm 1{,}988\sqrt{\frac{8{,}4^2}{50} + \frac{7{,}4^2}{40}}$$

$3{,}9 \pm 3{,}3(0{,}6$ a $7{,}2)$

14. a. H_0: $\mu_1 - \mu_2 = 0$
H_a: $\mu_1 - \mu_2 \neq 0$
b. 2,18
c. Entre 0,02 e 0,05
d. Rejeitar H_0: a média etária difere

16. a. H_0: $\mu_1 - \mu_2 = 0$
H_a: $\mu_1 - \mu_2 > 0$
b. 38
c. $t = 1{,}80$, $gl = 25$
Valor p entre 0,025 e 0,05
d. Rejeitar H_0; concluir que a pontuação média é mais elevada se os pais tiverem educação de nível superior.

18. a. H_0: $\mu_1 - \mu_2 \geq 120$
H_a: $\mu_1 - \mu_2 < 120$
b. −2,10
Entre 0,01 e 0,025
c. 32 a 118
d. Maior tamanho de amostra

19. a. 1, 2, 0, 0, 2
b. $\bar{d} = \Sigma d_i/n = 5/5 = 1$
c. $$s_d = \sqrt{\frac{\Sigma(d_i - \bar{d})^2}{n - 1}} = \sqrt{\frac{4}{5 - 1}} = 1$$
d. $$t = \frac{\bar{d} - \mu}{s_d/\sqrt{n}} = \frac{1 - 0}{1/\sqrt{5}} = 2{,}24$$
$gl = n - 1 = 4$
Valor p entre 0,025 e 0,05
Valor $p = 0{,}05$; rejeitar H_0

20. a. 3, −1, 3, 5, 3, 0, 1
b. 2
c. 2,08
d. 2
e. 0,07 a 3,93

21. H_0: $\mu_d = 0$
H_a: $\mu_d > 0$
$\bar{d} = 0{,}625$
$s_d = 1{,}30$
$$t = \frac{\bar{d} - \mu_d}{s_d/\sqrt{n}} = \frac{0{,}625 - 0}{1{,}30/\sqrt{8}} = 1{,}36$$
$gl = n - 1 = 7$
Valor p entre 0,10 e 0,20
Valor $p > 0{,}05$; não rejeitar H_0

22. 0,16 a 0,35

24. $t = 1{,}63$
Valor p entre 0,10 e 0,20
Não rejeitar H_0

26. a. $t = -0{,}60$
Valor p maior que 0,40
Não rejeitar H_0
b. −0,103
c. 0,39; maior tamanho de amostra

27. a. $\bar{\bar{x}} = (30 + 45 + 36)/3 = 37$

$$\text{SSTR} = \sum_{j=1}^{k} n_j(\bar{x}_j - \bar{\bar{x}})^2$$
$$= 5(30 - 37)^2 + 5(45 - 37)^2 + 5(36 - 37)^2$$
$$= 570$$
$$\text{MSTR} = \frac{\text{SSTR}}{k - 1} = \frac{570}{2} = 285$$

b. $$\text{SSE} = \sum_{j=1}^{k} (n_j - 1)s_j^2$$
$$= 4(6) + 4(4) + 4(6{,}5) = 66$$
$$\text{MSE} = \frac{\text{SSE}}{n_T - k} = \frac{66}{15 - 3} = 5{,}5$$

c. $$F = \frac{\text{MSTR}}{\text{MSE}} = \frac{285}{5.5} = 51{,}82$$
Da tabela F (numerador com 2 graus de liberdade e denominador 12), o valor p é menor que 0,01.
Uma vez que o valor $p = \alpha = 0{,}5$, rejeitamos a hipótese nula de que as médias das três populações sejam iguais.

d.

Fonte de Variação	Soma dos Quadrados	Graus de Liberdade	Média Quadrática	F
Tratamentos	570	2	285	51,82
Erro	66	12	5,5	
Total	636	14		

28. a. MSTR = 268
b. MSE = 92
c. Não se pode rejeitar H_0 porque o valor p é maior que 0,10
d.

Fonte de Variação	Soma dos Quadrados	Graus de Liberdade	Média Quadrática	F
Tratamentos	536	2	268	2,91
Erro	828	9	92	
Total	1.364	11		

30. **a.** 1.200; 3
300; 12
$F = 16$
b. Rejeitar H_0 porque o valor p é menor que 0,01

32.

	Fabricante 1	Fabricante 2	Fabricante 3
Média amostral	23	28	21
Variância da amostra	6,67	4,67	3,33

$\bar{\bar{x}} = (23 + 28 + 21)/3 = 24$

$SSTR = \sum_{j=1}^{k} n_j(\bar{x}_j - \bar{\bar{x}})^2$

$= 4(23 - 24)^2 + 4(28 - 24)^2 + 4(21 - 24)^2 = 104$

$MSTR = \frac{SSTR}{k-1} = \frac{104}{2} = 52$

$SSE = \sum_{j=1}^{k} (n_j - 1)s_j^2$

$= 3(6{,}67) + 3(4{,}67) + 3(3{,}33) = 44{,}01$

$MSE = \frac{SSE}{n_T - k} = \frac{44.01}{12 - 3} = 4{,}89$

$F = \frac{MSTR}{MSE} = \frac{52}{4{,}89} = 10{,}63$

Da tabela F (numerador com 2 graus de liberdade e 9 no denominador), o valor p é menor que 0,01
Uma vez que o valor $p = \alpha = 0{,}05$, rejeitamos a hipótese nula de que o tempo médio necessário para misturar um lote de matérias seja o mesmo para cada fabricante.

34. Médias amostrais: 81, 79, 88; $F = 4{,}99$
O valor p está entre 0,025 e 0,05
Diferença significativa; Vale do Silício

36. Há diferenças significativas; $F = 3{,}70$
O valor p está entre 0,025 e 0,05

38. 8.934 a 11.066

40. **a.** H_0: $\mu_1 - \mu_2 = 0$
H_a: $\mu_1 - \mu_2 > 0$
b. $t = 0{,}60$, $gl = 57$
Valor p maior que 20
Não rejeitar H_0

42. **a.** 15 (ou US$ 15.000)
b. 9,81 a 20,19
c. 11,5%

44. Médias amostrais: 58,6; 48,8; 60,1; $F = 18{,}59$
Valor $p \approx 0$; diferença significativa

46. Médias amostrais: 7,41; 6,11; 7,06; $F = 9{,}33$
Valor $p < 0{,}01$; diferença significativa

Capítulo 11

2. **a.** $\bar{p} = \frac{n_1\bar{p}_1 + n_2\bar{p}_2}{n_1 + n_2} = \frac{200(0{,}22) + 300(0{,}16)}{200 + 300} = 0{,}1840$

$z = \frac{\bar{p}_1 - \bar{p}_2}{\sqrt{\bar{p}(1-\bar{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$

$= \frac{0{,}22 - 0{,}16}{\sqrt{0{,}1840(1 - 0{,}1840)\left(\frac{1}{200} + \frac{1}{300}\right)}} = 1{,}70$

Valor $p = 0{,}5000 - 0{,}4554 = 0{,}0446$
b. Valor $p = 0{,}05$; rejeitar H_0

3. $\bar{p}_1 = 220/400 = 0{,}55$ $\bar{p}_2 = 192/400 = 0{,}48$

$\bar{p}_1 - \bar{p}_2 \pm z_{0,025}\sqrt{\frac{\bar{p}_1(1-\bar{p}_1)}{n_1} + \frac{\bar{p}_2(1-\bar{p}_2)}{n_2}}$

$0{,}55 - 0{,}48 \pm 1{,}96\sqrt{\frac{0{,}55(1-0{,}55)}{400} + \frac{0{,}48(1-0{,}48)}{400}}$

$0{,}07 \pm 0{,}0691$ (0,0009 a 0,1391)
Mais de 7% dos executivos prevêem um aumento dos empregos de tempo integral; o intervalo de confiança mostra que a diferença pode ser de 0% a 14%.

4. **a.** 0,46; 0,28
b. 0,18
c. 0,0777
d. 0,1023 a 0,2577; maior proporção de republicanos

6. **a.** 0,803
b. 0,849
c. H_0: $p_1 - p_2 \geq 0$
H_a: $p_1 - p_2 < 0$
d. Valor $p = 0{,}0104$
Rejeitar H_0

8. **a.** H_0: $p_1 - p_2 = 0$
H_a: $p_1 - p_2 \neq 0$
b. 0,13
c. 0,0404; a conclusão é de que existe diferença.
d. Sim; atrair o grupo etário mais jovem

10. Valor $p = 0{,}0322$
Rejeitar H_0

11. **a.** Freqüências esperadas: $e_1 = 200(0{,}40) = 80$
$e_2 = 200(0{,}40) = 80$
$e_3 = 200(0{,}20) = 40$
Freqüências reais: $f_1 = 60, f_2 = 120, f_3 = 20$

$\chi^2 = \frac{(60-80)^2}{80} + \frac{(120-80)^2}{80} + \frac{(20-40)^2}{40}$

$= \frac{400}{80} + \frac{1600}{80} + \frac{400}{40}$

$= 5 + 20 + 10 = 35$

Graus de liberdade: $k - 1 = 2$
$\chi^2 = 35$ mostra que o valor $p \approx 0$
Valor $p = 0{,}01$; rejeitar H_0
b. Rejeitar H_0 se $\chi^2 \geq 9{,}210$
$\chi^2 = 35$; rejeitar H_0

12. $\chi^2 = 15{,}33$; $gl = 3$
Valor p menor que 0,005
Rejeitar H_0

13. H_0: $p_{ABC} = 0{,}29$; $p_{CBS} = 0{,}28$; $p_{NBC} = 0{,}25$; $p_{IND} = 0{,}18$
H_a: As proporções não são
$p_{ABC} = 0{,}29$; $p_{CBS} = 0{,}28$; $p_{NBC} = 0{,}25$; $p_{IND} = 0{,}18$
Freqüências esperadas: $300(0{,}29) = 87$, $300(0{,}28) = 84$

$300(0{,}25) = 75$, $300(0{,}18) = 54$
$e_1 = 87$, $e_2 = 84$, $e_3 = 75$, $e_4 = 54$
Freqüências reais: $f_1 = 95, f_2 = 70, f_3 = 89, f_4 = 46 = 6{,}87$
Graus de liberdade: $k - 1 = 3$
$\chi^2 = 6{,}87$, o valor p está entre 0,05 e 0,10
Não rejeitar H_0

14. $\chi^2 = 29{,}51$; $gl = 5$
Valor $p \approx 0$
Rejeitar H_0

16. **a.** $\chi^2 = 12{,}21$; $gl = 3$
O valor p está entre 0,005 e 0,01
A conclusão é que há uma diferença para 2003
b. 21%, 30%, 15%, 34%
Maior utilização do cartão de débito
c. 51%

18. $\chi^2 = 16{,}31$; $gl = 3$
Valor p menor que 0,005
Rejeitar H_0

19. H_0: A variável coluna é independente da variável linha
H_a: A variável coluna não é independente da variável linha
Freqüências esperadas

	A	B	C
P	28,5	39,9	45,6
Q	21,5	30,1	34,4

$$\chi^2 = \frac{(20 - 28{,}5)^2}{28{,}5} + \frac{(44 - 39{,}9)^2}{39{,}9} + \frac{(50 - 46{,}5)^2}{45{,}6} + \frac{(30 - 21{,}5)^2}{21{,}5} + \frac{(26 - 30{,}1)^2}{30{,}1} + \frac{(30 - 34{,}4)^2}{34{,}4}$$

$= 7{,}86$

Graus de liberdade: $(2 - 1)(3 - 1) = 2$
$\chi^2 = 7{,}86$, o valor p está entre 0,01 e 0,025
Rejeitar H_0

20. $\chi^2 = 19{,}77$; $gl = 4$
valor de p menor que 0,005
Rejeitar H_0

21. H_0: O tipo de passagem comprada depende do tipo de vôo
H_a: O tipo de passagem comprada não independe do tipo de vôo
Freqüências esperadas:
$e_{11} = 35{,}59$ $e_{12} = 15{,}41$
$e_{21} = 150{,}73$ $e_{22} = 65{,}27$
$e_{31} = 455{,}68$ $e_{32} = 197{,}32$

Passagem	Vôo	Freqüência Observada (f_i)	Freqüência Esperada (e_i)	$(f_i - e_i)^2/e_i$
Primeira classe	Doméstico	29	35,59	1,22
Primeira classe	Internacional	22	15,41	2,82
Business/Executiva	Doméstico	95	150,73	20,61
Business/Executiva	Internacional	121	65,27	47,59
Full-fare	Doméstico	518	455,68	8,52
Full-fare	Internacional	135	197,32	19,68
Totais		920	$\chi^2 =$	100,43

Graus de liberdade: $(3 - 1)(2 - 1) = 2$
$\chi^2 = 100{,}43$, valor $p \approx 0$
Rejeitar H_0

22. **a.** $\chi^2 = 7{,}36$; $gl = 2$
O valor p está entre 0,25 e 0,05
Rejeitar H_0
b. Domésticos 47,2%

24. **a.** $\chi^2 = 10{,}60$; $gl = 4$
O valor p está entre 0,025 e 0,05
Rejeitar H_0; não-independente
b. Efeito negativo mais acentuado sobre a obtenção do diploma à medida que as horas aumentam

26. **a.** $\chi^2 = 7{,}85$; $gl = 3$
O valor p está entre 0,025 e 0,05
Rejeitar H_0
b. Produtos farmacêuticos, 98,6%

28. **a.** H_0: $p_1 - p_2 = 0$
H_a: $p_1 - p_2 \neq 0$
b. 0,31; 0,26
c. $z = 2{,}04$; valor $p = 0{,}0414$
Rejeitar H_0; a conclusão é que há diferença
d. 0,0475; 0,0025 a 0,0975

30. $z = 2{,}37$; valor $p = 0{,}0178$
Rejeitar H_0

32. **a.** 0,16
b. H_0: $p_1 - p_2 = 0$
H_a: $p_1 - p_2 > 0$
c. $z = 3{,}49$; valor $p = 0$
Rejeitar H_0

34. $\chi^2 = 4{,}64$; $gl = 2$
O valor p está entre 0,05 e 0,10
Não rejeitar H_0

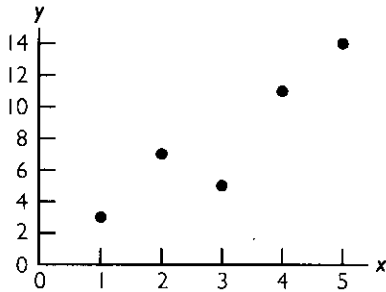
36. $\chi^2 = 42{,}53$; $gl = 4$
Valor $p \approx 0$; rejeitar H_0

38. $\chi^2 = 23{,}37$; $gl = 3$
Valor $p \approx 0$; rejeitar H_0

40. **a.** $\chi^2 = 12{,}86$; $gl = 2$
Valor p menor que 0,005
Rejeitar H_0
b. 66,9; 30,3; 2,9
54,0; 42,0; 4,0

42. **a.** 24,01; 41,16; 20,46; 8,37
O último lançamento (entrada) combina 3 e 4
b. $\chi^2 = 6{,}17$; $gl = 3$
Valor p maior do que 0,10
Não rejeitar H_0: binomial

Capítulo 12

1. a.

b. Parece haver uma relação linear entre x e y

c. Muitas linhas retas diferentes podem ser traçadas para prover uma aproximação linear à relação entre x e y; no item (d), determinaremos a equação de uma linha reta que "melhor" representa a relação de acordo com o critério dos mínimos quadrados.

d. Somatórios necessários para calcular a inclinação e a interseção com o eixo y:

$\Sigma x_i = 15,\quad \Sigma y_i = 40,\quad \Sigma(x_i - \bar{x})(y_i - \bar{y}) = 26,$

$\Sigma(x_i - \bar{x})^2 = 10$

$$b_1 = \frac{\Sigma(x_i - \bar{x})(y_i - \bar{y})}{\Sigma(x_i - \bar{x})^2} = \frac{26}{10} = 2{,}6$$

$b_0 = \bar{y} - b_1\bar{x} = 8 - (2{,}6)(3) = 0{,}2$

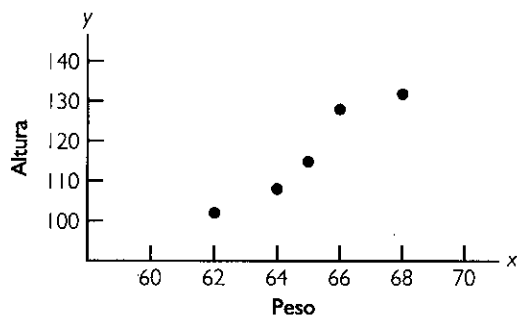
$\hat{y} = 0{,}2 - 2{,}6x$

e. $\hat{y} = 0{,}2 + 2{,}6x = 0{,}2 + 2{,}6(4) = 10{,}6$

2. b. Parece haver uma relação linear entre x e y

d. $\hat{y} = 30{,}33 - 1{,}88x$

e. 19,05

4. a.

b. Indica que pode haver uma relação linear entre altura e peso.

c. Muitas linhas retas diferentes podem ser traçadas para prover uma aproximação linear à relação entre altura e peso; no item (d) determinaremos a equação de uma linha reta que "melhor" representa a relação de acordo com o critério dos mínimos quadrados.

d. Somatórios necessários para calcular a inclinação e a interseção com o eixo y:

$\Sigma x_i = 325,\quad \Sigma y_i = 585,\quad \Sigma(x_i - \bar{x})(y_i - \bar{y}) = 110,$

$\Sigma(x_i - \bar{x})^2 = 20$

$$b_1 = \frac{\Sigma(x_i - \bar{x})(y_i - \bar{y})}{\Sigma(x_i - \bar{x})^2} = \frac{110}{20} = 5{,}5$$

$b_0 = \bar{y} - b_1\bar{x} = 117 - (5.5)(65) = -240{,}5$

$\hat{y} = -240{,}5 + 5{,}5x$

e. $\hat{y} = -240{,}5 + 5{,}5(63) = 106$

A estimativa de peso é 106 libras (48 kg)

6. c. $\hat{y} = -10{,}16 + 0{,}18x$

e. 11,95, ou aproximadamente US$ 12.000

8. c. $\hat{y} = 490{,}21 + 204{,}24x$

d. US$ 1.307

10. b. $\hat{y} = 51{,}82 + 0{,}145x$

c. 84,4

12. c. $\hat{y} = 1.293 + 0{,}3165x$

d. 25.031

14. b. $\hat{y} = 28{,}30 - 0{,}0415x$

c. 26,2

15. a. $\hat{y}_i = 0{,}2 + 2{,}6x_i$ e $\bar{y} = 8$

x_i	y_i	$\hat{y}_i$	$y_i - \hat{y}_i$	$(y_i - \hat{y}_i)^2$	$y_i - \bar{y}$	$(y_i - \bar{y})^2$
1	3	2,8	0,2	0,04	–5	25
2	7	5,4	1,6	2,56	–1	1
3	5	8,0	–3,0	9,00	–3	9
4	11	10,6	0,4	0,16	3	9
5	14	13,2	0,8	0,64	6	36
				SSE = 12,40		SST = 80

SSR = SST − SSE = 80 − 12,4 = 67,6

b. $r^2 = \frac{\text{SSR}}{\text{SST}} = \frac{67{,}6}{80} = 0{,}845$

A reta dos mínimos quadrados proporcionou um bom ajuste; 84,5% da variabilidade em y foi explicada pela reta dos mínimos quadrados.

c. $r = \sqrt{0{,}845} = +0{,}9192$

16. a. SSE = 6,3325, SST = 114,80, SSR = 108,47

b. $r^2 = 0{,}945$

c. $r = -0{,}9721$

18. a. A equação de regressão estimada e a média da variável dependente:

$\hat{y} = 1.790{,}5 + 581{,}1x,\quad \bar{y} = 3.650$

A soma dos quadrados dos erros e a soma total dos quadrados:

SSE = $\Sigma(y_i - \hat{y}_i)^2$ = 85.135,14

SSE = $\Sigma(y_i - y_i)^2$ = 335.000

Desse modo, SSR = SST – SSE

= 335.000 – 85.135,14 = 249.864,86

b. $r^2 = \frac{\text{SSR}}{\text{SST}} = \frac{24.864{,}86}{335.000} = 0{,}746$

A reta dos mínimos quadrados é responsável por 74,6% da soma total dos quadrados.

c. $r = \sqrt{0{,}746} = +0{,}8637$

20. a. $\hat{y} = -48{,}11 + 2{,}3325x$

b. $r^2 = 0{,}82$

c. US$ 173.500

22. **a.** $\hat{y} = -745{,}80 + 117{,}917x$
b. $r^2 = 0{,}7071$
c. $r = +0{,}84$

23. **a.** $s^2 = \text{MSE} = \dfrac{\text{SSE}}{n-2} = \dfrac{12{,}4}{3} = 4{,}133$
b. $s = \sqrt{\text{MSE}} = \sqrt{4{,}133} = 2{,}033$
c. $\Sigma(x_i - \bar{x})^2 = 10$

$$s_{b_1} = \frac{s}{\sqrt{\Sigma(x_i - \bar{x})^2}} = \frac{2{,}033}{\sqrt{10}} = 0{,}643$$

d. $t = \dfrac{b_1 - \beta_1}{s_{b_1}} = \dfrac{2{,}6 - 0}{0{,}643} = 4{,}04$

Da tabela t (3 graus de liberdade), a área na cauda superior está entre 0,01 e 0,025
O valor p está entre 0,02 e 0,05
Valor p real = 0,0272
Uma vez que o valor $p = \alpha$, rejeitamos H_0: $b_1 = 0$

e. $\text{MSR} = \dfrac{\text{SSR}}{1} = 67{,}6$

$$F = \frac{\text{MSR}}{\text{MSE}} = \frac{67{,}6}{4{,}133} = 16{,}36$$

Da tabela F (1 grau de liberdade no numerador e 3 no denominador), o valor p está entre 0,025 e 0,05
Valor p real = 0,0272
Uma vez que o valor $p = \alpha$, rejeitamos H_0: $\beta_1 = 0$

Fonte de Variação	Soma dos Quadrados	Graus de Liberdade	Média Quadrática	F
Regressão	67,6	1	67,6	16,36
Erro	12,4	3	4,133	
Total	80	4		

24. **a.** 2,11
b. 1,453
c. 0,262
d. Significativa; o valor p é menor que 0,01
e. Significativa; o valor p é menor que 0,01

26. **a.** $s^2 = \text{MSE} = \dfrac{\text{SSE}}{n-2} = \dfrac{85.135{,}14}{4} = 21.283{,}79$

$s = \sqrt{\text{MSE}} = \sqrt{21.283{,}79} = 145{,}89$

$\Sigma(x_i - \bar{x})^2 = 0{,}74$

$$s_{b_1} = \frac{s}{\sqrt{\Sigma(x_i - \bar{x})^2}} = \frac{145{,}89}{\sqrt{0{,}74}} = 169{,}59$$

$$t = \frac{b_1 - \beta_1}{s_{b_1}} = \frac{581{,}08 - 0}{169{,}59} = 3{,}43$$

Da tabela t (4 graus de liberdade), a área na cauda está entre 0,01 e 0,025
O valor p está entre 0,02 e 0,05
Valor p real = 0,0266
Uma vez que o valor $p = \alpha$, rejeitamos H_0: $\beta_1 = 0$

b. $\text{MSR} = \dfrac{\text{SSR}}{1} = \dfrac{249.864{,}86}{1} = 249.864{,}86$

$$F = \frac{\text{MSR}}{\text{MSE}} = \frac{249.864{,}86}{21.283{,}79} = 11{,}74$$

Da tabela F (1 grau de liberdade no numerador e 4 no denominador), o valor p está entre 0,025 e 0,05
Valor real de $p = 0{,}0266$
Uma vez que o valor $p = \alpha$, rejeitamos H_0: $\beta_1 = 0$

Fonte de Variação	Soma dos Quadrados	Graus de Liberdade	Média Quadrática	F
Regressão	29.864,86	1	29.864,86	11,74
Erro	85.135,14	4	21.283,79	
Total	335.000	5		

28. Elas estão relacionadas; o valor p é menor que 0,01

30. Está relacionado significativamente; o valor p é menor que 0,01

32. **a.** $s = 2{,}033$

$\bar{x} = 3$, $\Sigma(x_i - \bar{x})^2 = 10$

$$s_{\hat{y}_p} = s\sqrt{\frac{1}{n} + \frac{(x_p - \bar{x})^2}{\Sigma(x_i - \bar{x})^2}}$$

$$= 2.033\sqrt{\frac{1}{5} + \frac{(4-3)^2}{10}} = 1{,}11$$

b. $\hat{y} = 0{,}2 + 2{,}6a = 0{,}2 + 2{,}6(4) = 10{,}6$

$\hat{y}_p \pm t_{\alpha/2} s_{\hat{y}_p}$
$10{,}6 \pm 3{,}182(1{,}11)$
$10{,}6 \pm 3{,}53$, ou 7,07 a 14,13

c. $s_{\text{ind}} = s\sqrt{1 + \dfrac{1}{n} + \dfrac{(x_p - \bar{x})^2}{\Sigma(x_i - \bar{x})^2}}$

$$= 2{,}033\sqrt{1 + \frac{1}{5} + \frac{(4-3)^2}{10}} = 2{,}32$$

d. $\hat{y} \pm t_{\alpha/2} s_{\text{ind}}$
$10{,}6 \pm 3{,}182(2{,}32)$
$10{,}6 \pm 7{,}38$, ou 3,22 a 17,98

34. Intervalo de confiança: –0,4 a 4,98
Intervalo de previsão: –2,27 a 7,31

35. **a.** $s = 145{,}89$, $\bar{x} = 3{,}2$, $\Sigma(x_i - \bar{x})^2 = 0{,}74$

$\hat{y} = 1790{,}5 + 581{,}1x = 1790{,}5 + 581{,}1(3) = 3533{,}8$

$$s_{\hat{y}_p} = s\sqrt{\frac{1}{n} + \frac{(x_p - \bar{x})^2}{\Sigma(x_i - \bar{x})^2}}$$

$$= 145{,}89\sqrt{\frac{1}{6} + \frac{(3-3{,}2)^2}{0{,}74}} = 68{,}54$$

$\hat{y}_p \pm t_{\alpha/2} s_{\hat{y}_p}$
$3.533{,}8 \pm 2{,}776(68{,}54)$
$3.533{,}8 \pm 190{,}27$, ou US\$ 3.343,53 a US\$ 3.724,07

b. $s_{\text{ind}} = s\sqrt{1 + \dfrac{1}{n} + \dfrac{(x_p - \bar{x})^2}{\Sigma(x_i - \bar{x})^2}}$

$$= 145{,}89\sqrt{1 + \frac{1}{6} + \frac{(3-3{,}2)^2}{0{,}74}} = 161{,}19$$

$\hat{y}_p \pm t_{\alpha/2} s_{\text{ind}}$
$3.533{,}8 \pm 2{,}776(161{,}19)$
$3.533{,}8 \pm 447{,}46$, ou US\$ 3.086,34 a US\$ 3.981,26

36. **a.** 80.86
b. 78,58 a 83,14
c. 72,92 a 88,80

38. **a.** US$ 5.046,67
b. US$ 3.815,10 a US$ 6.278,24
c. Não fora da reta

40. **a.** 9
b. $\hat{y} = 20,0 + 7,21x$
c. 1,3626
d. SSE = SST – SSR = 51.984,1 – 41.587,3 = 10.396,8
MSE = 10.396,8/7 = 1.485,3

$$F = \frac{\text{MSR}}{\text{MSE}} = \frac{41,587.3}{1485.3} = 28.0$$

Da tabela F (1 grau de liberdade no numerador e 7 no denominador), o valor p é menor que 0,01
Valor p real = 0,0011
Uma vez que o valor $p = \alpha = 0,05$, rejeitamos H_0: $\beta_1 = 0$
e. $\hat{y} = 20,0 + 7,21(50) = 380,5$, ou US$ 380.500

42. **a.** $\hat{y} = 80,0 + 50,0x$
b. 30
c. Significativa; o valor p é menor que 0,01
d. US$ 680.000

44. **b.** Sim
c. $\hat{y} + 37,1 - 0,779x$
d. Significativa. Valor p = 0,003
e. $r^2 = 0,434$. não há um bom ajuste
f. US$ 12,27 a US$ 22,90
g. US$ 17,47 a US$ 39,05

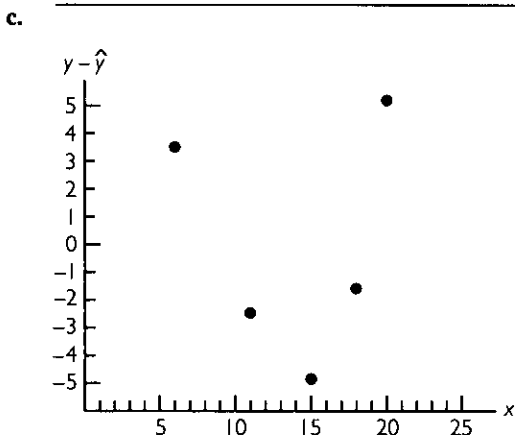
45. **a.** $\Sigma x_i = 70$, $\Sigma y_i = 76$, $\Sigma(x_i - \bar{x})(y_i - \bar{y}) = 200$, $\Sigma(x_i - \bar{x})^2 = 126$

$$b_1 = \frac{\Sigma(x_i - \bar{x})(y_i - \bar{y})}{\Sigma(x_i - \bar{x})^2} = \frac{200}{126} = 1.5873$$

$b_0 = \hat{y} - b_1\bar{x} = 15,2 - (1,5873)(14) = 7,0222$
$\hat{y} = -7,02 + 1,59x$

b.

x_i	y_i	$\hat{y}_i$	$y_i - \hat{y}_i$
6	6	2,52	3,48
11	8	10,47	22,47
15	12	16,83	24,83
18	20	21,60	21,60
20	30	24,78	5,22

c.

Com somente cinco observações é difícil de determinar se as hipóteses (suposições) são satisfeitas; entretanto, o diagrama de dispersão unidimensional (*dot plot*) sugere uma curvatura nos resíduos, algo que indicaria que as hipóteses do termo de erro não são satisfeitas; o diagrama de dispersão desses dados também indica que a relação subjacente entre x e y pode ser curvilínea.

46. **a.** $\hat{y} = 2,32 + 0,64x$
b. Não, a variância não parece ser idêntica para todos os valores de x

47. **a.** Admitamos que x = despesas de publicidade e y = receita
$\hat{y} = 29,4 + 1,55x$
b. SST = 1002, SSE = 310,28, SSR = 691,72

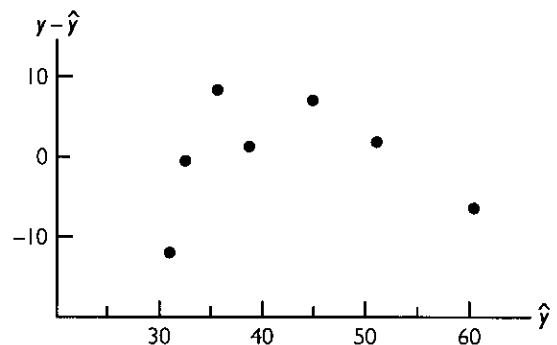
$$\text{MSR} = \frac{\text{SSR}}{1} = 691,72$$

$$\text{MSE} = \frac{\text{SSE}}{n-2} = \frac{310,28}{5} = 62,0554$$

$$F = \frac{\text{MSR}}{\text{MSE}} = \frac{691,72}{62,0554} = 11,15$$

Da tabela F (1 grau de liberdade no numerador e 5 no denominador), o valor p está entre 0,01 e 0,025
Valor p real = 0,0206
Uma vez que o valor $p = \alpha = 0,05$, concluímos que as duas variáveis são relacionadas

c.

x_i	y_i	$\hat{y}_i = 29,40 + 1,55x_i$	$y_i - \hat{y}_i$
1	19	30,95	−11,95
2	32	32,50	−0,50
4	44	35,60	8,40
6	40	38,70	1,30
10	52	44,90	7,10
14	53	51,10	1,90
20	54	60,40	−6,40

d. A plotagem residual nos leva a questionar a suposição de uma relação linear entre x e y; não obstante a relação ser significativa ao nível $\alpha = 0,05$, seria extremamente perigoso extrapolar além do intervalo dos dados.

48. **b.** Sim

50. **a.** $\hat{y} = 9,26 + 0,711x$
b. Significativa; valor p = 0,001
c. $r^2 = 0,744$; bom ajuste
d. US$ 13,53

52. **a.** Títulos com beta = 0,95
b. Significativa; valor $p = 0{,}029$
c. $r^2 = 0{,}470$; não há um bom ajuste
d. A Texas Instruments tem um risco maior

54. **a.** $\hat{y} = 10{,}5 + 0{,}953x$
b. Relação significativa; valor $p = 0{,}000$
c. US$ 2.874 a US$ 4.952
d. Sim

56. **a.** Relação linear negativa
b. $\hat{y} = 8{,}10 - 0{,}344x$
c. Significativa; valor $p = 0{,}002$
d. $r^2 = 0{,}711$; ajuste razoavelmente bom
e. 5,2 a 7,6 dias

58. **a.** $\hat{y} = 5{,}85 + 0{,}830x$
b. Significativa; valor $p = 0{,}000$
c. 84,65 pontos
d. 65,35 a 103,96

Capítulo 13

2. **a.** A equação de regressão estimada é
$\hat{y} = 45{,}06 + 1{,}94x_1$
Uma estimativa de y quando $x_1 = 45$ é
$\hat{y} = 45{,}06 + 1{,}94(45) = 132{,}36$
b. A equação de regressão estimada é
$\hat{y} = 85{,}22 + 4{,}32x_2$
Uma estimativa de y quando $x_2 = 15$ é
$\hat{y} = 85{,}22 + 4{,}32(15) = 150{,}02$
c. A equação de regressão estimada é
$\hat{y} = -18{,}37 + 2{,}01x_1 + 4{,}7x_2$
Uma estimativa de y quando $x_1 = 45$ e $x_2 = 15$ é
$\hat{y} = -18{,}37 + 2{,}01(45) + 4{,}74(15) = 143{,}18$

4. **a.** US$ 255.000

5. **a.** A saída de dados (*output*) do Minitab é apresentada na Figura D13.5a
b. A saída de dados do Minitab é apresentada na Figura D13.5b

Figura D13.5a

```
The regression equation is
Revenue = 88.6 + 1.60 TVAdv

Predictor       Coef    SE Coef       T       p
Constant      88.638      1.582   56.02   0.000
TVAdv         1.6039     0.4778    3.36   0.015

S = 1.215     R-sq = 65.3%     R-sq(adj) = 59.5%

Analysis of Variance

SOURCE            DF       SS       MS       F       p
Regression         1   16.640   16.640   11.27   0.015
Residual Error     6    8.860    1.477
Total              7   25.500
```

Figura D13.5b

```
The regression equation is
Revenue = 83.2 + 2.29 TVAdv + 1.30 NewsAdv

Predictor       Coef    SE Coef       T       p
Constant      83.230      1.574   52.88   0.000
TVAdv         2.2902     0.3041    7.53   0.001
NewsAdv       1.3010     0.3207    4.06   0.010

S = 0.6426     R-sq = 91.9%     R-sq(adj) = 88.7%

Analysis of Variance

SOURCE            DF       SS       MS       F       p
Regression         2   23.435   11.718   28.38   0.002
Residual Error     5    2.065    0.413
Total              7   25.500
```

c. É 1,60 no item (a) e 2,29 no item (b); no item (a) o coeficiente é uma estimativa da alteração da receita em razão da mudança de uma unidade nos gastos com propaganda de televisão; no item (b), representa uma estimativa da alteração da receita em virtude da mudança de uma unidade nos gastos com propaganda de televisão quando a quantidade da propaganda em jornais se mantém constante.

d. Receita = 83,2 + 2,29(3,5) + 1,30(1,8) = 93,56, ou US$ 93.560

6. **a.** PPG = 0,354 + 0,000888 HR
b. PPG = 0,865 − 0,0837 MRR
c. PPG = 0,709 + 0,00140 HR − 0,103 MMR
d. 54,9%

8. **a.** Retorno (rentabilidade) = 247 − 32,8 Segurança + 34,6 taxa de despesa (ExpRatio)
b. 70,2

10. **a.** PPG = −1,22 + 3,96 FG%
b. Um aumento de 0,01 na porcentagem de *field goals* (FG%) aumentará a PPG em aproximadamente 0,04
c. PPG = −1,23 + 4,82 FG% − 2,59 %3Pt Adv + 0,344 Turnover Adv
e. 0,6432

12. **a.** $R^2 = \frac{\text{SSR}}{\text{SST}} = \frac{14.052,2}{15.182,9} = 0,926$

b. $R_a^2 = 1 - (1 - R^2)\frac{n-1}{n-p-1}$

$= 1 - (1 - 0,926)\frac{10-1}{10-2-1} = 0,905$

c. Sim; depois de ajustar o número de variáveis independentes no modelo, vemos que 90,5% da variabilidade em y foi a responsável.

14. **a.** 0,75
b. 0,68

15. **a.** $R^2 = \frac{\text{SSR}}{\text{SST}} = \frac{23,435}{25,5} = 0,919$

$R_a^2 = 1 - (1 - R^2)\frac{n-1}{n-p-1}$

$= 1 - (1 - 0,919)\frac{8-1}{8-2-1} = 0,887$

b. A análise de regressão múltipla é preferível porque tanto R^2 e R_a^2 mostram que uma porcentagem aumentada da variabilidade de y é explicada quando ambas as variáveis independentes são usadas.

16. **a.** Não, $R^2 = 0,153$
b. Melhor ajuste com regressão múltipla.

18. **a.** $R^2 = 0,564$, $R_a^2 = 0,511$
b. O ajuste não é muito bom

19. **a.** $\text{MSR} = \frac{\text{SSR}}{p} = \frac{6216,375}{2} = 3108,188$

$\text{MSE} = \frac{\text{SSE}}{n-p-1} = \frac{507,75}{10-2-1} = 72,536$

b. $F = \frac{\text{MSR}}{\text{MSE}} = \frac{3108,188}{72,536} = 42,85$

Da tabela F (2 graus de liberdade no numerador e 7 no denominador), o valor p é menor que 0,01
Uma vez que o valor $p = \alpha$, o modelo global é significativo

c. $t = \frac{b_1}{s_{b_1}} = \frac{0,5906}{0,0813} = 7,26$

O valor p é menor que 0,01
Uma vez que o valor $p = \alpha$, β_1 é significativa.

d. $t = \frac{b_2}{s_{b_2}} = \frac{0,4980}{0,0567} = 8,78$

O valor p é menor que 0,01
Uma vez que o valor $p = \alpha$, β_2 é significativa

20. **a.** Significativa; valor $p = 0,000$
b. Significativa; valor $p = 0,000$
c. Significativa; valor $p = 0,002$

22. **a.** SSE = 4.000, $s^2 = 571,43$, MSR = 6.000
b. Significativa; o valor p é menor que 0,01

23. **a.** $F = 28,38$
Valor $p = 0,002$
Uma vez que o valor $p = \alpha$, há uma relação significativa
b. $t = 7,53$
Valor $p = 0,001$
Uma vez que o valor $p = \alpha$, β_1 é significativa e x_1 não deve ser eliminado do modelo
c. $t = 4,06$
Valor $p = 0,10$
Uma vez que o valor $p = \alpha$, β_2 é significativa e x_2 não deve ser eliminado do modelo

24. **a.** Rejeitar H_0: $\beta_1 = \beta_2 = 0$; valor $p = 0,000$
b. HR: Rejeitar H_0: $\beta_1 = 0$; valor $p = 0,000$
MMR: Rejeitar H_0: $\beta_2 = 0$; valor $p = 0,000$

26. **a.** Significativa; valor $p = 0,000$
b. Todas as variáveis independentes são significativas

28. **a.** Com o Minitab, o intervalo de confiança de 95% é 132,16 a 154,15
b. Com o Minitab, o intervalo de previsão de 95% é 111,15 a 175,17

29. **a.** Veja a saída de dados do Minitab na Figura D13.5b
$\hat{y} = 83,230 + 2,2902(3,5) + 1,3010(1,8) = 93,588$, ou US$ 93.588
b. Resultados do Minitab: 92,840 a 94,335, ou US$ 92.840 a US$ 94.335
c. Resultados do Minitab: 91,774 a 95.401, ou US$ 91.774 a US$ 95.401

30. **a.** 58,37% a 75,03%
b. 35,24% a 90,59%

32. **a.** $E(y) = \beta_0 + \beta_1 x_1 + b_2 x_2$

em que $x_2 = \begin{cases} 0 \text{ se corresponder ao nível 1} \\ 1 \text{ se corresponder ao nível 2} \end{cases}$

b. $E(y) = \beta_0 + \beta_1 x_1 + \beta_2(0) = \beta_0 + \beta_1 x_1$

c. $E(y) = \beta_0 + \beta_1 x_1 + \beta_2(1) = \beta_0 + \beta_1 x_1 + \beta_2$

d. $\beta_2 = E(y \mid \text{nível } 2) - E(y \mid \text{nível } 1)$

β_1 é a alteração em $E(y)$ para uma mudança de 1 unidade em x_1 ao manter-se x_2 constante

34. **a.** US\$ 15.300, porque $\beta_3 = 15{,}3$

b. $\hat{y} = 10{,}1 - 4{,}2(2) + 6{,}8(8) + 15{,}3(0) = 56{,}1$

Previsão de vendas: US\$ US\$ 56.100

c. $\hat{y} = 10{,}1 - 4{,}2(1) + 6{,}8(3) + 15{,}3(1) = 41{,}6$

Previsão de vendas: US\$ 41.600

36. **a.** $\hat{y}$ = 1,86 + 0,291 Meses + 1,10 Tipo – 0,609 Técnico

b. Significativa; valor $p = 0{,}002$

c. A adição do técnico não é significativa; valor $p = 0{,}167$

38. **a.** $\hat{y}$ = –91,8 + 1,08 Idade + 252 Pressão Arterial + 8,74 Fumante

b. Significativo; valor $p = 0{,}01$

c. O intervalo de previsão de 95% é 21,35 a 47,18, ou uma probabilidade de 0,2135 a 0,4718; parar de fumar e iniciar algum tipo de tratamento para reduzir o nível de pressão arterial

40. **b.** 67,39

42. **a.** $\hat{y} = -1{,}41 + 0{,}0235x_1 + 0{,}00486x_2$

b. Significativa

c. $R^2 = 0{,}937$; $R^2_a = 9{,}19$; bom ajuste

d. Ambas significativas

44. **a.** Pontuação = 50,6 + 1,56 Resistência à Recessão

b. $r^2 = 0{,}431$; não há um bom ajuste

c. Pontuação = 33,5 + 1,90 Resistência à Recessão + 2,61 Acessibilidade

Significativa

$R^2_2 = 0{,}784$; ajuste muito melhor

46. **a.** MPG Cidade = 24,1 – 2,10 Cilindradas

Significativa; valor $p = 0{,}000$

b. MPG Cidade = 26,4 – 2,44 Cilindradas – 1,20 T4R

c. Significativa; valor $p = 0{,}016$

d. MPG Cidade = 33,3 – 4,15 Cilindradas – 1,24 Tração4 + 2,16 OitoCil

e. Significativa global e individualmente

Índice Remissivo

A

A Companhia Colgate-Palmolive, estatística e, 21,
Abordagem pelo valor crítico, teste de hipóteses e
    teste bicaudal, 320-323
    teste unicaudal, 315-320
Alliance Data Systems, regressão linear simples e a, 427
Amostra(s), 12, 237-238. *Veja também* Estimação por intervalo
Amostra, coeficiente de correlação da, 102
Amostra, covariância da, 98, 99
Amostra, determinação do tamanho (estimação por intervalo) da, 287-288, 291-292
Amostra, espaço da, 130
Amostra, estatística(s) da, 72, 244-245
Amostra, média da, 239
    fórmula, 72-73
    para dados agrupados, 108
Amostra, ponto de, 131
Amostragem
    MeadWestvaco Corporation, exemplo da, 237
    problema exemplo, 239
    Veja também Estimação por ponto; Distribuições de amostragem; Métodos de amostragem; Amostragem aleatória simples
Amostragem aleatória estratificada, 262-263
Amostragem aleatória simples
    capacidade do Excel para, 268-269
    capacidade do Minitab para, 268
    de população finita, 240-241
    de população infinita, 241
    independente, 354
Amostragem com substituição, 241
Amostragem de conveniência, 264
Amostragem intencional, 264
Amostragem por conglomerado, 263
Amostragem sem substituição, 241
Amostragem sistemática, 263-264
Amostras aleatórias simples independentes, 354
Amostras dependentes. *Veja* Comparações envolvendo médias
Amplitude, 81
Amplitude interquartil (AIQ)
    definição, 82
    fórmula, 82
Análise de variância (ANOVA), 372-373
    capacidade do Excel para, 396-397
    capacidade do Minitab para, 396
    estimativa da variância da população entre tratamentos, 377-378
    estimativa da variância da população dentro dos tratamentos, 378
    hipóteses para, 374
    resultados de computador para, 381-382
    tabela ANOVA, 381
    teste da igualdade das médias de k populações, 376
        teste F (comparação de estimativas da variância), 378-380
    Veja também Comparações envolvendo médias
    visão conceitual, 374-376
Análise exploratória de dados
    apresentação de ramo-e-folha, 38-41
    desenhos esquemáticos (*box plots*), 94-95
    regra de cinco itens, 94
Análise residual (regressão linear simples)
    plotagem residual em relação a x, 467-468
    plotagem residual em relação a $\hat{y}$, 468-469
    resíduo da observação i, 466
Apresentações tabular e gráfica
    análise exploratória de dados (gráfico de ramo-e-folha), 38-41
    capacidade do Excel para, 62-70
    capacidade do Minitab para, 60-61
    Colgate-Palmolive Company, exemplo de uso de, 21
    dados qualitativos, resumo de
    dados quantitativos, resumo de distribuição de freqüência, 28-30
    diagrama de dispersão, 46-48
    distribuição de freqüência, 23
    distribuições cumulativas, 32-33
    distribuições de freqüência relativa e de freqüência percentual, 24
    distribuições de freqüência relativa e de freqüência percentual, 30-31
    gráfico de dispersão unidimensional (*dot plot*), 31
    gráficos em barras e em setores, 24-25
    histograma, 31-32
    linha de tendência, 46-48
    ogiva, 34
    paradoxo de Simpson, 45-46
    tabulações cruzadas, 43-45
Aproximação pela normal de probabilidades binomiais, 223-224
Área como medida da probabilidade, 207-209
Arredondamento de erros, 91
Assimetria, 87-88
    população, estimação por intervalo e, 284
Associação entre duas variáveis. *Veja* Medidas numéricas

B

Bayes, Thomas, 157
Bernoulli, Jakob, 182

Bolsa de valores, 5
Business Week, estatística e a, 1
Butler, Marty, 309

C

Censo, 12
Citibank, distribuição de probabilidade discreta e o, 169
Classes
amplitude das, 29
com extremidade aberta, 34-35
limites de, 29, 34
limites superior e inferior, 30
na distribuição de freqüência, 25, 28-30
número de, 29
ponto médio, 30, 38-39
Classes de extremidade aberta, 34-35
Clemance, Philip, 427
Coeficiente de confiança, 275
Coeficiente de correlação, 443-444
Coeficiente de correlação do momento produto de Pearson
dados da amostra, 102
dados da população, 102
Coeficiente de correlação do momento produto de Pearson
dados da população, 102
dados da amostra, 102
interpretação do, 102-103
Coeficiente de determinação, 440-444
Coeficiente de determinação múltiplo, 497-499
Coeficiente de determinação múltiplo ajustado, 498
Coeficiente de variação, 84
Combinações. *Veja* Probabilidade
Comparações envolvendo médias
análise de variância (ANOVA)
capacidade do Excel para, 395
capacidade do Excel para, 396-397
capacidade do Minitab para, 396
comparação de estimativas da variância: teste F, 378-380
conselho prático, 358
duas populações: σ_1 e σ_2 conhecidos
estimação por intervalo de $\mu_1 - \mu_2$, 354-356
estimativa de variância populacional entre tratamentos, 377-378
estimativa dentro dos tratamentos da variância da população, 378
exemplo da Fisons Corporation, 353
inferências sobre a diferença entre as médias de
introdução à, 372-376
resultados de computador para, 381-382
tabela ANOVA, 381
teste da igualdade das médias de k populações, 376-382
testes de hipóteses sobre $(\mu_1 - \mu_2)$, 356-358
inferências sobre a diferença entre as médias de duas populações: amostras relacionadas, 368-370
capacidade do Excel para, 396-397
capacidade do Minitab para, 396
inferências sobre a diferença entre as médias de duas populações: σ_1 e σ_2 desconhecidos
capacidade do Excel para, 396
capacidade do Minitab para, 394
conselho prático, 363
estimação por intervalo de $m_1 - m_2$, 360-361
testes de hipótese sobre $(m_1 - m_2)$, 389-391
Comparações envolvendo proporções e teste de independência
eficiência de ajuste, 407-409
capacidade do Excel para, 424, 426
capacidade do Minitab para, 424
exemplo da United Way, 399
inferências sobre a diferença entre duas proporções de população
capacidade do Minitab para, 423
estimação por intervalo de $p_1 - p_2$, 400-402
teste de hipótese sobre $(p_1 - p_2)$, 402-403
teste de independência, 411-415
capacidade do Excel para, 424-426
capacidade do Minitab para, 424
teste de hipótese para proporções de população multinomial, 406-409
Complemento de A, 143
Complemento de um evento, 143
Computadores
análise estatística e, 13-14
regressão linear simples e, 462-463
Veja também Excel; Minitab
Conglomerados, 263
Conjunto de dados, 4
Contabilidade, estatística na, 3
Contagens-*z*, 88
Controle de qualidade, gráficos de barras no, 24
Costeletas, 94
Covariância, 98-99
Covariância da população, 99
Cunningham, Keith, 271

D

Dados
bimodal, 75
contínuo, 7
definição, 4
elementos, 4
multimodal, 75
observações, 4
qualitativo, 6
quantitativo, 6
secção transversal, de, 6
série histórica, 6
validade dos, verificação da, 91
variáveis, 5
Dados agrupados, 107
média da amostra para, 108
média da população para, 109
variância da amostra para, 108-109
variância da população para, 109

Dados da amostra, Coeficiente de correlação do momento produto de Pearson e, 102
Dados da população, coeficiente de correlação do produto-momento de Pearson e, 102
Dados de secção transversal, 6
Dados de séries temporais, 6
Dados multimodais, 75
Dados qualitativos, 6
- moda como medida de posição para, 75
- resumos
  - distribuição de freqüência, 23
  - distribuições de freqüência relativa e de freqüência percentual, 24
  - gráficos em barras e em setores, 24-25

Dados quantitativos, 6
- discretos, 7
- resumos
  - distribuição de freqüência, 28-30
  - distribuições cumulativas, 32-33
  - distribuições de freqüência relativa e de freqüência percentual, 30-31
  - gráfico de dispersão unidimensional (dot plot), 31
  - histograma, 31-32
  - ogivas, 34

Dados quantitativos contínuos, 7
Dados quantitativos discretos, 7
Definição de variável aleatória contínua, 171
Desenhos esquemáticos (*box plots*), 94-95
Desvio em torno da média, 82
Desvio padrão, 84
- de p, 258-259
- de x, 249-250
- definição, 84
- fórmula, 84

Diagrama em árvore, 132-133
Diagrama de ramo-e-folha, 38-41
- capacidade do Minitab para, 60-61

Diagrama de ramo-e-folha alongado, 39-40
Diagrama de Pareto, 25
Diagrama de Venn, 143
Diagramas de dispersão, 46-48
- capacidade do Excel para, 65-67
- capacidade do Minitab para, 61
- regressão linear simples e, 431-432
- regressão múltipla e, 490-491

Distribuição amostral de $\bar{p}$, 257
- desvio padrão e, 258-259
- formato da, 259
- valor esperado e, 258
- valor prático da, 259-260

Distribuição amostral de $\bar{x}$
- definição, 246
- desvio padrão e, 249-250
- formato da, 250-251
- para o problema exemplo, 252
- tamanho da amostra e, 253-255
- valor esperado e, 249
- valor prático da, 252-253

Distribuição contínua exponencial de probabilidade, 228
Distribuição de freqüência
- capacidade do Excel para, 62-63
- classes na, 25, 28-30
- dados qualitativos e, 23
- dados quantitativos e, 28-30
- definição, 23
- soma de freqüências, 25

Distribuição de freqüência relativa
- cumulativa, 33
- dados qualitativos e, 24
- dados quantitativos e, 30-31

Distribuição "studentizada" de amplitudes, valores críticos para, 563-564
Distribuição de Poisson de probabilidade, distribuições exponenciais e, 228. *Veja também* Distribuições discretas de probabilidade
Distribuição de probabilidade, 173
Distribuição de probabilidade binomial, 181-182. *Veja também* Distribuições de probabilidades discretas
Distribuição de quiquadrado, 408, 539-540
Distribuição exponencial de probabilidade. *Veja* Distribuições contínuas de probabilidade
Distribuição F, 541-542
Distribuição hipergeométrica de probabilidade, 195-196
Distribuição normal de probabilidade. *Veja* Distribuições contínuas de probabilidade
Distribuição normal padrão, 535
Distribuição normal padrão de probabilidade, 212-213
Distribuição percentual acumulada de freqüência, 33
Distribuição percentual de freqüência
- acumulativa, 33
- dados qualitativos e, 24
- dados quantitativos e, 30-31

Distribuição relativa acumulada de freqüência, 33
- definição, 278
- distribuição t, 536-538
- estimação por intervalo e, 278-279
- tabela, 280

Distribuição uniforme discreta de probabilidade, 174
Distribuição uniforme de probabilidade. *Veja* Distribuições contínuas de probabilidade
Distribuições contínuas de probabilidade
- aproximação pela normal de probabilidades binomiais, 223-224
- área como medida de probabilidade, 207-209
- capacidade do Excel para, 235-236
- capacidade do Minitab para, 234-235
- distribuição exponencial de probabilidade
  - cálculo de probabilidades para, 226-228
  - função de Poisson e, 228
  - função exponencial de densidade de probabilidade, 226
  - probabilidades cumulativas, 227
- distribuição normal de probabilidade
  - cálculo de probabilidades para, 218
  - curva normal, 211-212

distribuição normal padrão de probabilidade, 212-218
exemplo da companhia de pneus, 218-220
função densidade normal de probabilidade, 211
função densidade normal padrão de probabilidade, 213
distribuição uniforme de probabilidade, 207
função densidade uniforme de probabilidade, 207
exemplo da Procter & Gamble, 205
Distribuições cumulativas, dados quantitativos e, 32-33
Distribuições de amostragem, 246-248
distribuição de probabilidade binomial, 181-182
Distribuições de probabilidade discretas, 173, 175
experimento binomial, 181-183
função probabilidade binomial, 186
problema da loja de roupas, 183-187
tabelas para, 187-188
valor esperado e variância de, 188-189
capacidade do Excel para, 202-203
capacidade do Minitab para, 202
distribuição de probabilidade hipergeométrica, 195-196
distribuição de probabilidade uniforme discreta, 174
distribuição probabilidade de Poisson, 191-192
exemplo de intervalos de tempo, 192-193
intervalos de comprimento ou de distância, 193
exemplo do Citibank, 169
função probabilidade discreta, 173
valor esperado, 177
variância, 178
variáveis aleatórias, 170
variáveis aleatórias contínuas, 171
variáveis aleatórias discretas, 170-171

E

Elementos, 4
Equação de regressão, 429-430
Excel e, 484
regressão múltipla e, 488
Equação de regressão estimada, 429-430
para estimativa e previsão (regressão múltipla), 507-508
declive e intercepto *y* para, 471
Equação de regressão múltipla estimada, 433-434
Erro padrão
da estimativa, 449
da média, 250
da proporção, 259
Erro quadrático médio (MSE), 449
Erros na obtenção de dados, 10
Erros tipo I, 313-314
Erros tipo II, 313-314
Escala de medição do intervalo, 5-6
Escala de medida de relação, 6
Escala nominal de medição, 5
Escala ordinal de medida 5
Escalas de medida. *Veja* Medidas, escalas de
Estatística
aplicações em administração e economia, 2-4
definição, 2-3
Estatística da regressão, Excel e a, 486
Estatística de teste
para eficiência do ajuste, 406-407
para igualdade de k médias da população, 378
para independência, 413
para teste de hipóteses sobre $(\mu_1 - \mu_2)$: (μ_1) e (μ_2) conhecidos, 356-357
para teste de hipóteses sobre $(\mu_1 - \mu_2)$: (μ_1) e (μ_2) desconhecidos, 362
para teste de hipóteses sobre a média da população: σ desconhecido, 328-329
para teste de hipóteses sobre a proporção populacional, 336
teste de hipóteses e, 316-317
Estatística descritiva, 10-12. *Veja também* Medidas numéricas; Apresentações tabular e gráfica
Estatística na economia, 3-4
Estatística na produção, 3
Estimação por intervalo
capacidade do Excel para, 305-307
capacidade do Minitab para, 303-304
determinação do tamanho da amostra, 287-288
envolvendo proporções e um teste de independência
exemplo da Food Lion, 271
média da população: σ conhecido
conselho prático, 276
definição, 276
margem de erro e, 273-276
média da população: Û desconhecido, 278
amostra pequena, 282-284
conselho prático, 282
distribuição t e, 278-280
margem de erro e, 279, 304-305
proporção da população, 290-291
determinação do tamanho da amostra, 291-292
regressão linear simples e, 456
resumo de procedimentos, 284
teste de hipóteses e, 324-325
Veja também Comparações envolvendo médias; Comparações
Estimação por intervalo, objetivo da, 272
Estimador agrupado, 402
Estimador por pontos,
da diferença entre as médias de duas populações, 354
da diferença entre proporções de duas populações, 400
definição, 72, 272
Estimativa da variância da população dentro dos tratamentos, 378
Estimativa de σ^2 dentro dos tratamentos, 375-376
Estimativa de tratamentos agrupados SS, 375
Estimativa entre tratamentos, 375, 377
Estimativa por ponto, 244-245
regressão linear simples e, 456
Estratos, 262-263
Estudos com observação (não-experimentais), 10
Estudos estatísticos
como ajuda na tomada de decisão, 10
experimentais, 8,
não-experimentais (de observação), 10

Estudos experimentais, 7
Estudos não-experimentais (de observações), 10
Eventos
independente, 151, 152
mutuamente exclusivo, 146, 152
probabilidades e, 139-141
Veja também Complemento de um evento
Eventos independentes, 151
lei da multiplicação para, 151-152
Excel
amostragem aleatória simples, 268-269
análise de regressão, 484-486
análise de variância (ANOVA), 396-397
apresentações tabular e gráfica, 62-70
distribuições contínuas de probabilidade, 235
distribuições discretas de probabilidade, 202-203
estimação por intervalo, 305-307
inferência sobre a diferença entre duas médias de população: σ_1 e σ_2 desconhecidos, 396
inferências sobre a diferença entre duas médias de população: amostras relacionadas, 396
inferências sobre a diferença entre duas médias de população: σ_1 e σ_2 conhecidos, 395
medidas numéricas, 124-127
regressão múltipla, 526-528
teste de eficiência de ajuste, 424-425
teste de hipóteses, 347-351
teste de independência, 425-426
Experimento (de probabilidade), 130-131
aleatório, 137
Experimentos aleatórios, 137
Experimentos em múltiplas etapas, 131-133

F

Fator, 373
Fator de correção de continuidade, 223
Fator de correção de população finita, 250
Fatorial, 134
Finanças, estatística em, 3
Fisons Corporation, estudos estatísticos de população e, 353
Folha, 39
Fontes de dados
erros na obtenção de dados, 10
estudos estatísticos, 8, 10
existentes, 7-8,
Food Lion, estimação por intervalo e, 271
Forma da distribuição, 87Fowle, William R., 21
Freqüência relativa, 24
Função de densidade uniforme de probabilidade, 207
Função de Poisson de probabilidade, 192
Função de probabilidade, 173
Função de probabilidade binomial, 183
Função densidade de probabilidade, 207
Função densidade normal padrão, 213
Função discreta de probabilidade, 173
Função exponencial de densidade de probabilidade, 226
Função hipergeométrica de probabilidade, 195-196
Função normal de densidade da probabilidade, 211

G

Galton, Francis, 428
Gauss, Carl Friedrich, 432
Gosset, William Sealy, 278
Gráfico(s) de barras, 10,11
capacidade do Excel para, 64-65
dados qualitativos e, 24-25
objetivo do, 34-35
Gráficos de dispersão unidimensional (*dot plot*)
capacidade do Minitab para, 60
dados quantitativos e, 31
Gráficos de resíduos, 467-469
Gráficos de setores, dados qualitativos e, 24-25
Grau de convicção, 135-136
Graus de liberdade, 278
Griggs, Bill, 487

H

Harkey, Bobby, 271
Haskell, Michael, 129
Hipótese alternativa, 310-312. *Veja também* Teste de hipóteses
Hipótese nula
definição, 310
desenvolvimento, 310-312
Veja também Teste de hipóteses
Hipóteses
na regressão múltipla, 500-501
na regressão linear simples, 447-448, 466-469
Histograma, 10, 11
capacidade do Excel para, 64-65, 66
capacidade do Minitab para, 60
dados quantitativos e, 31-32
propósito do, 34
simétrico, 32

I

i-ésimo resíduo, 440
Inferência estatística, 12-13, 238
International Paper, uso da regressão múltipla pela, 487
Internet como fonte de dados, 8
Interseção de A e B, 144
Interseção de dois eventos, 144
Intervalo de confiança, 274
para β_1, 450-451
para o valor médio de y, 456-457
regressão linear simples e, 456
teste de hipóteses e, 325
Intervalo de previsão, 456
para um valor individual de *y*, 458-460

Intervalos de distância, distribuição de Poisson de probabilidade e, 193
Intervalos de distância, distribuição de probabilidade de Poisson e, 193
Intervalos de tempo, função de Poisson de probabilidade e, 209-192-193

K

Kahn, Joel, 205
Karter, Stacey, 169

L

Lei da adição, 143-146
Lei da multiplicação, 151-152
Levantamento por amostragem, 12
Limite inferior de classe, 29
Limite superior de classe, 29
Linha de tendência, 46-48

M

Margem de erro, 272. *Veja também* Estimação por intervalo
Margem de lucro bruto, 5
Marketing, estatística em, 3
McCarthy, John A., 71
MeadWestvaco Corporation, amostragem e a, 237
Média, 72-73
  ajustada, 77
Média ajustada, 77
Média da população
  fórmula, 73
  para dados agrupados, 109
  Veja também Comparações envolvendo médias; Teste de hipóteses; Estimação por intervalo
Média geral da amostra, 375-376
Média ponderada, 106-107
Mediana, 74, 77
Médias. *Veja* Comparações envolvendo médias
Médias de *k* populações. *Veja* Comparações envolvendo médias
Medidas numéricas
  análise exploratória de dados
    desenhos esquemáticos (*box plots*), 94-95
    regra de cinco itens, 94
  associação entre duas variáveis, medidas do coeficiente de
  capacidade do Excel para, 124-127
  capacidade do Minitab para, 122-124
  correlação, 102-104
    covariância, 98-101
  dados agrupados, 107-109
  forma da distribuição, medidas do, 87-88
  média ponderada, 106-107
  pontos fora da curva, detecção de, 90-91
  posição relativa, medidas de
    teorema de Chebyshev, 89, 90-91
    regra empírica, 90
    contagens-z, 88
  posição, medidas de
    média, 72-73, 77
    mediana, 74, 77
    moda, 74-75
    percentis, 75-76
    quartis, 76-77
  Small Fry Designs, exemplo da, 71
  variabilidade, medidas de, 81
    amplitude interquartil, 82
    amplitude, 81
    coeficiente de variação, 84
    desvio padrão, 84
    variância, 82-84
Medidas, escalas de
  com intervalo, 5-6
  nominal, 5
  ordenada, 5
  de proporção, 6
Método clássico de atribuição de probabilidades, 135, 141
Método de atribuição de probabilidades pela freqüência relativa, 135
Método dos mínimos quadrados
  regressão linear simples e, 431-434, 444
  regressão múltipla e, 489-493
Método subjetivo de atribuição de probabilidades, 136
Métodos de amostragem, 262
  amostragem por conglomerado, 263
  amostragem de conveniência, 264
  amostragem por julgamento, 264
  amostragem aleatória estratificada, 262-263
  amostragem sistemática, 263-264
Minitab
  amostragem aleatória simples, 268
  análise de regressão, 483-484
  análise de variância (Anova), 396
  apresentações tabular e gráfica, 60-61
  distribuições contínuas de probabilidade, 234-235
  distribuições discretas de probabilidade, 202
  estimação por intervalo, 303-304
  inferências sobre a diferença entre duas médias de população: amostras relacionadas, 395-396
  inferências sobre a diferença entre duas médias de população: σ_1 e σ_2 desconhecidos, 394
  inferências sobre a diferença entre duas proporções de população, 423
  medidas numéricas, 122-124
  regressão linear simples, 462-463
  regressão múltipla, 525
  teste de hipóteses, 345-347
  teste de independência, 425
  teste de qualidade do ajuste, 424
Moda, 74-75
Modelo de regressão, 466
  regressão múltipla e, 535
Morton International, probabilidades e a, 129

Multicolinearidade, 504-505
Myerson, Roger, 210

N

Nível de confiança, 274
Nível de significância, 313-314
  observados, 319
Nível de significância observado, 319
Notação, 565-566
Notação de somatório, 565-566
Notações taquigráficas, 566
Números aleatórios
  gerados via computador,
  tabela de, 240

O

Observações, 5, 7
Ogivas, 34
Operações aritméticas, 6

P

p-ésimo percentil
  cálculo do, 75
  definição, 75
Paradoxo de Simpson, 45-46
Parâmetros. *Veja* Regressão múltipla; Parâmetros da população
Parâmetros da população
  amostragem e, 238-239
  definição, 72
Pareto, Vilfredo, 25
Partição, 381
Pearson, Karl, 428
Percentis, 75-76
Permutações. *Veja* Probabilidade
Pesquisa com entrevista pessoal, 9
Pesquisas, 10
Pontos fora da curva
  detecção, 87
  exemplo, 94
População
  definição, 12, 238
  Veja também População finita; População infinita
População finita, amostragem aleatória simples e, 240-241
População infinita, amostragem aleatória simples e, 241
População multinomial. *Veja* Comparações envolvendo
  proporções e teste de independência
Posição central
  média e, 74, 77
  mediana e, 74, 77
Posição relativa
  contagens-*z*, 88
  regra empírica, 90
  teorema de Chebyshev, 89
Posição, medidas de, *Veja* Medidas numéricas
Probabilidade
  anterior, 155
  área como medida de, 207-209
  atribuição
    método clássico de 135, 140
    método da freqüência relativa para, 135
    método subjetivo para, 136
  combinações, 133
  condicional, 148-151
    eventos independentes, 151
    lei da multiplicação, 151-152
  definição, 130
  eventos e, 139-140
  exemplo da Morton International, 129
  experimentos, 130-131
  experimentos em múltiplas etapas, 142-145
  permutações, 134-135
  posterior, 155
  projeto exemplo da KP&L 136-137
  regras de contagem, 131-135
  relações básicas da
    complemento de um evento, 143
    lei da adição, 143-146
  teorema de Bayes, 155-158
Probabilidade *a priori*, 155
Probabilidade condicional, 148-151
  eventos independentes, 151
  lei da multiplicação, 151-152
Probabilidades associadas, 149
Probabilidades binomiais, 545-550
Probabilidades de Poisson, 552-557
Probabilidades marginais, 149
Probabilidades *a posteriori*, 155
Procter & Gamble e distribuição contínua de probabilidade, 205
Projeto de amostra independente, 368
Projeto de amostras relacionadas, 368
Proporção da população. *Veja* Comparações envolvendo
  proporções e um teste de independência; Teste de hipóteses;
  Estimação por intervalo

Q

Quadrado médio devido à regressão (MSR), 451-452
Quadrado médio devido ao erro (MSE), 378
Quadrado médio devido aos tratamentos (MSTR), 377-378
Quartis, 76-77

R

Ramo, 18-19
Regra empírica, 90
Regra de cinco itens, 94
Regras de contagem. *Veja* Probabilidade
Regras de rejeição (testes de hipóteses)
  para teste de cauda inferior: abordagem do valor crítico, 319
  uso do valor *p*, 318
Regressão da média dos quadrados (MSR), 451-452

Regressão linear simples
  Alliance Data Systems, exemplo da, 427
  análise residual e, 466
  capacidade do Excel para, 484-486
  capacidade do Minitab para, 462-463, 483
  coeficiente de determinação, 440-444
  equação de regressão, 429
  equação de regressão estimada, 429-430
  estimação por intervalo, 456
  estimação por ponto, 456
  intervalo de confiança para o valor médio de y, 462-463
  intervalo de previsão para um valor individual de y, 458-460
  método dos mínimos quadrados, 431-435
  modelo de regressão, 429
  plotagem residual em relação a x, 467-468
  plotagem residual em relação a ŷ, 468,-469
  solução de computador para, 504-505
  suposições do modelo, 447-448
  teste de significância
    estimativa de Û, 448449
    intervalo de confiança para ß1, 450-451
    precauções acerca do, 452-453
    teste F, 451-452
    teste t, 449-450
Regressão múltipla
  capacidade do Excel para, 526-528
  capacidade do Minitab para, 525
  coeficiente múltiplo de determinação, 497-499
  equação de regressão, 488
  equação de regressão múltipla estimada, 489
  estimação e previsão, 507
  exemplo da International Paper, 487
  método dos mínimos quadrados, 489-493
  modelo da, 488-489
  modelo de regressão, 488
  suposições do modelo, 500-501
  teste de significância,
    multicolinearidade, 504-505
    teste F, 502-503
    teste t, 504
  variáveis qualitativas independentes, 509-514
Relação preço/rendimento, 5
Requisitos básicos para a designação de probabilidades. *Veja* Probabilidade
Resumos gráficos, 10-12. *Veja também* Apresentações tabular e gráfica
Resumos numéricos, 10-12
Resumos tabulares, 10-12
Reta de regressão estimada, 430

S

Significância
  geral, 501
  individual, 501, 504
  níveis de, 313-314
  observada, 319
  Veja também Regressão linear múltipla; Regressão linear simples
Significância geral, 501
Significância individual, 501
Símbolo no painel eletrônico, 5
Small Fry Designs, estatística descritiva e, 71
Soma dos quadrados devido à regressão, 442-443
  regressão múltipla e, 497-498
Soma dos quadrados dos erros, 378, 440, 442-443
  regressão múltipla e, 497-498
Soma dos quadrados dos desvios, 432
Soma dos quadrados dos tratamentos, 377
Soma dos quadrados total, 442-443
  regressão múltipla e, 497-498
Somatórios duplos, 566
Superfície de resposta, 501

T

Tabela ANOVA, 381
  teste de significância na regressão e, 452
Tabela de contingência, 411
Tabela de probabilidade associada, 149-150
Tabulação cruzada, 43-44
  capacidade do Excel para, 67-70
  capacidade do Minitab para, 60-61
Técnica de amostragem não-probabilística, 264
Técnicas de amostragem de probabilidade, 264
Teorema de Bayes, 155-159
  abordagem tabular, 158
  análise da decisão e, 159
  fórmula, 157-158
Teorema de Chebyshev, 89-91
Teorema do limite central
  definição, 251
  demonstração teórica do, 255
Termo (ε) de erro, 447-448, 466
  na regressão múltipla, 500
Teste *F*
  estimativas da variância e, 378-380
  regressão linear simples e, 451-452
  regressão múltipla e, 502-503
Teste *t*
  regressão linear simples e, 449-450
  regressão múltipla e, 504
Teste da eficiência de ajuste, 406
  distribuição multinomial e, 409
Teste de Autocorrelação de Durbin-Watson, valores críticos para o, 559-561
Teste de hipóteses
  capacidade do Excel para, 347-351
  capacidade do Minitab para, 345-347
  erros tipo I e tipo II, 313-314
  exemplo da John Morrell & Company, 309
  média da população: σ conhecido
    estimação por intervalo e 324-325
    resumo, 323-324

teste bicaudal, 320-323
teste unicaudal, 315-320
média da população: σ desconhecido
resumo, 331-332
teste bicaudal, 330-331
teste unicaudal, 329-330
na tomada de decisão, 311
passos da, 323-324
pesquisa e, 310
proporção da população, 335-337
validade de uma afirmação, 311
Veja também Hipótese alternativa; Comparações envolvendo médias; Comparações envolvendo proporções e um teste de independência; Hipótese nula
Teste de hipóteses pela John Morrel & Company, 309
Teste de independência. *Veja* Comparações envolvendo proporções e um teste de independência
Teste de Mann-Whitney-Wilcoxon, valores de *TL* para, 562
Teste de tabela de contingência, 411
Testes bicaudais (teste de hipóteses)
média da população: σ conhecido, 320
critério do valor crítico, 322-323
critério do valor p, 321-322
resumo e aviso, 323-324
média da população: σ desconhecido, 330-331
resumo e aviso, 331-332
Testes de significância, 314
Testes unicaudais (teste de hipóteses)
média da população: σ conhecido, 315
critério do valor crítico, 319-320
critério do valor p, 317-318
estatística de teste, 316-317
resumo e aviso, 323-324
média da população: σ desconhecido, 328-330
resumo e aviso, 331-332
Tomada de decisão
dados e análise estatística e, 10-11
teste de hipóteses e, 311
Tratamentos, 373
Trentham, Charlene, 1
Tyler, Philip R., 399

U

União de A e B, 144
União de dois eventos, 144
Unidade de folha, 40
United Way, teste de independência do exemplo da, 399

V

Validade de uma afirmação, teste de hipóteses da, 311
Valor de mercado, 5
Valor esperado
de $\bar{p}$, 258
de uma variável aleatória discreta, 170
de $\bar{x}$, 249
para distribuição binomial, 188-189
valor p, teste de hipóteses e, 317-318, 321-322
Valor padronizado (contagem-z), 88
Valor planejado para σ, 288
Valores críticos
da distribuição de amplitude "studentizada", 563-564
para o Teste de Autocorrelação de Durbin-Watson, 559-561
Valores de (e^{-u}), 551
Valores T_L para o Teste de Mann-Whitney-Wilcoxon, 562
Variabilidade, medidas de. *Veja* Medidas numéricas
Variância, 82-83
de uma variável aleatória discreta, 177-178
para distribuição binomial, 188-189
Veja também Variância da população; Variância da amostra
Variância da amostra
fórmula, 82
para dados agrupados, 106-107
unidades elevadas ao quadrado, 82-83
variância de amostras em grupo, 364
Variância da população
fórmula, 82
para dados agrupados, 109
Variância de amostra agrupada, 364
Variável aleatória discreta
definição, 170
valor esperado de, 177
variância de, 178
Variável de resposta, 373, 501
Variável dependente, 428, 501
Variável independente, 428
Variável indicadora, 510
Variável qualitativa, 6
Variável quantitativa, 6
Variável simulada, 510
Variáveis
definição, 5
dependente, 373, 428, 501
escalas de medição para, 5-6
independente, 373, 428
qualitativa e quantitativa, 6
qualitativa independente, 509-514
resposta, 501
simulada (indicadora), 510
Variáveis aleatórias, 170
contínuas, 171
definição, 170
discretas, 170-171
Variáveis qualitativas independentes, 509-514

W

Williams, Marian, 487
Winkofsky, Edward P., 237

ESTATÍSTICA APLICADA
à Administração e Economia

2ª edição

Este livro proporciona uma efetiva introdução conceitual à estatística e suas aplicações, utilizando desenvolvimento metodológico fundamentado e notação adequada a cada tópico tratado. O único requisito matemático para seu estudo é o conhecimento da álgebra.

Orientado à análise de dados e de metodologia estatística, principal tônica do conteúdo, organiza-se em conjuntos de aplicações com discussão e desenvolvimento de cada técnica e resultados que fornecem subsídios para a solução dos problemas. *Estatística Aplicada à Administração e Economia* traz também uma bibliografia revista e atualizada como apêndice.

A obra apresenta ainda diversos exemplos, exercícios e estudos de caso e, na página do livro, no site da Thomson, aproximadamente 160 conjuntos de dados estão disponíveis para alunos e professores, tanto em formato Minitab como em formato Excel.

Aplicações

Livro-texto para a disciplina estatística nos cursos de Administração e Economia e para todos aqueles que utilizam ferramentas estatísticas nas áreas de Contabilidade, Finanças e Marketing, entre outras.

ISBN 13: 978-85-221-0521-2
ISBN 10: 85-221-0521-9